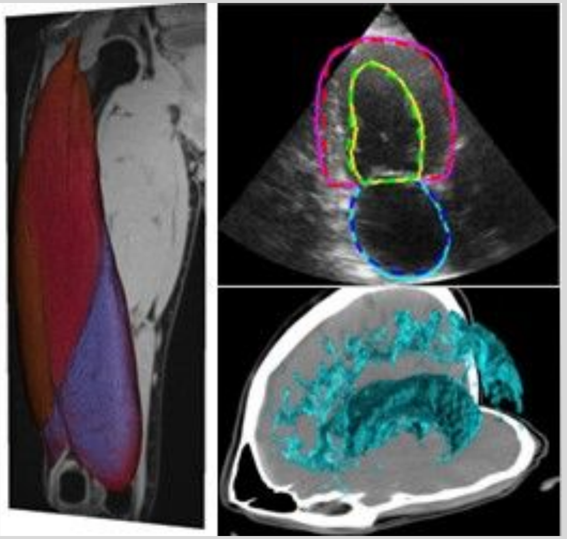
Deep learning for medical imaging school 2021

April 19—24 2021

Virtual edition



LABEX PRIMES
UNIVERSITÉ DE LYON

ÉTS
Le génie pour l'industrie

Weakly Supervised CNN Segmentation: Models and Optimization

Ismail Ben Ayed
Jose Dolz

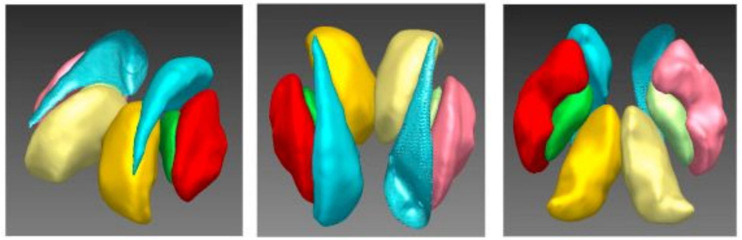
Why weak supervision is interesting?

Deep CNNs are dominating computer vision

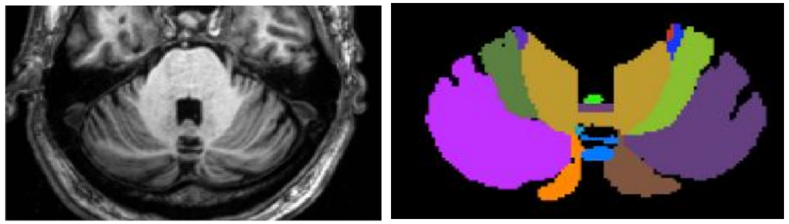
e.g., semantic segmentation

Cordts et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding," CVPR 2016

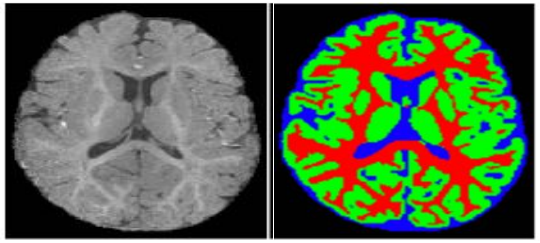
… and medical image analysis



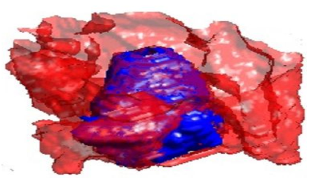
Subcortical structures
(Dolz et al., Neuroimage 2018)

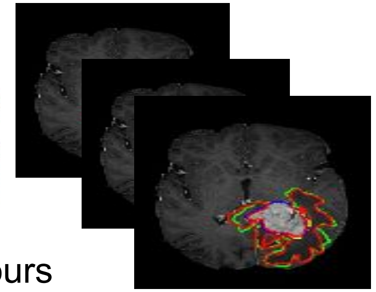


Cerebellum parcellation
(Carass et al., Neuroimage 2018)

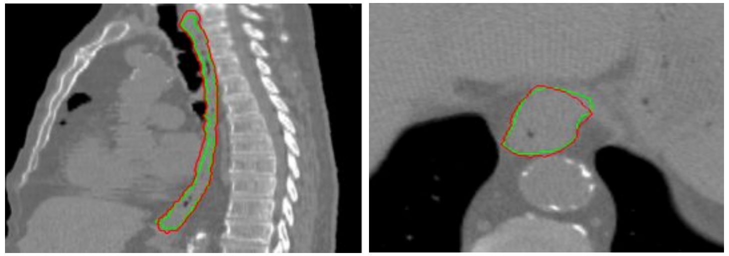


Brain tissues (6-month infant)
(Li et al., TMI 2019)

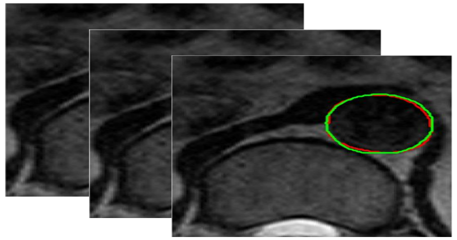


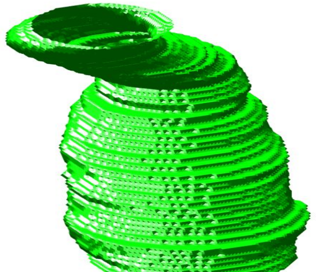


Brain tumours
(Njeh et al., CMIG 2015)



Organs at risk
(Dolz et al., Med. Phys. 2017)





Incidental findings
(Ben Ayed et al., MICCAI 2014)

But, massive and dense annotations are not always available

Full supervision

- more than 1h per image (even several hours for a medical image)

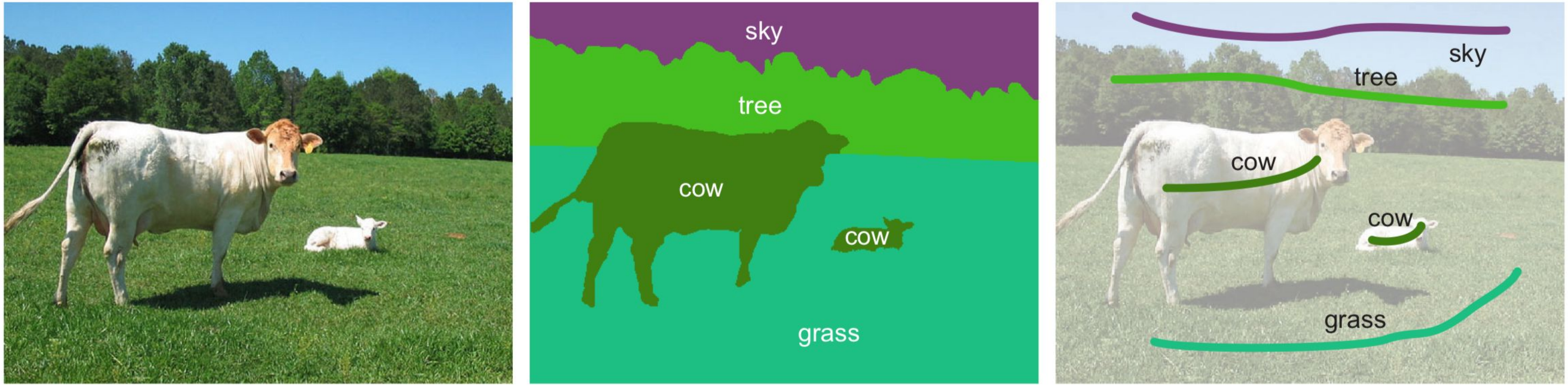
- Bottleneck for learning at large scale

Weak supervision
(e.g., image-level tags)

- 1s per label per image
- Scalable for large numbers of labels

person
horse
background

Semi-supervision with a lot of **non-annotated** data, and a **fraction** of points annotated



Full annotations

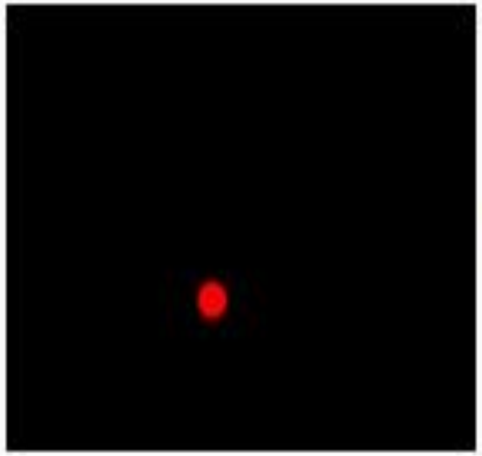
Semi-supervised

Figures from Lin et al. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, CVPR 2016

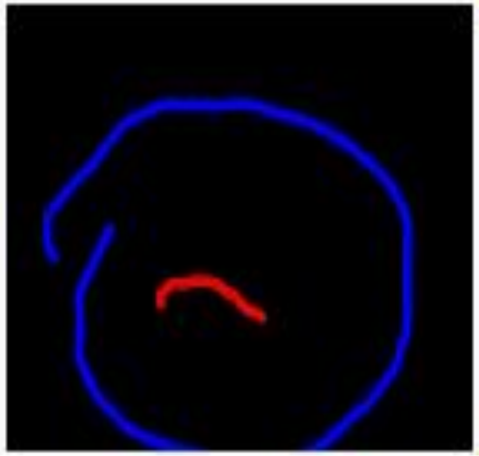
Forms of semi/weak supervision: Examples in segmentation

Car
Parking
Sky
No person

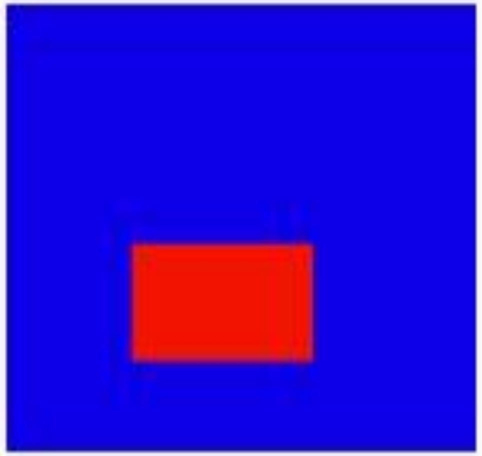
Image tags



points



scribbles

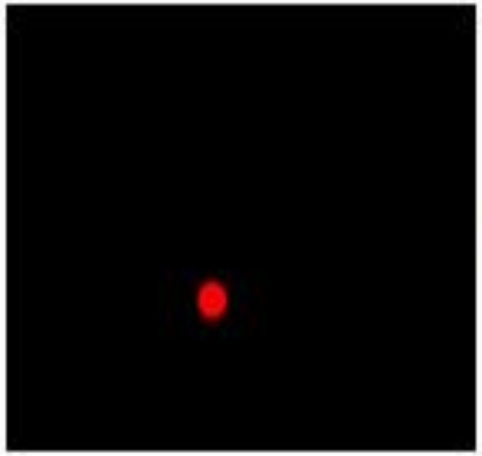


boxes

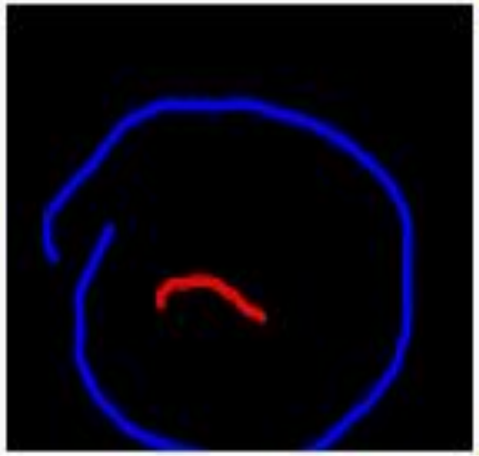
Forms of semi/weak supervision: Examples in segmentation

Car
Parking
Sky
No person

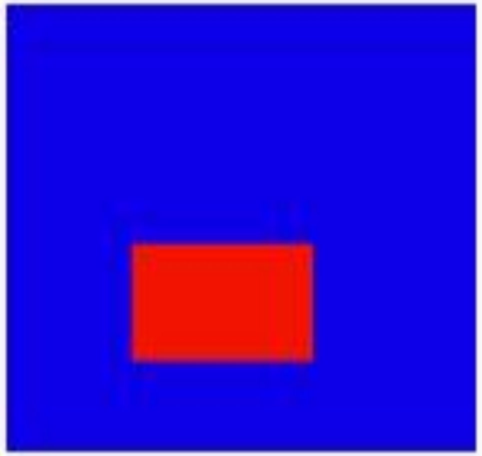
Image tags



points



scribbles



boxes

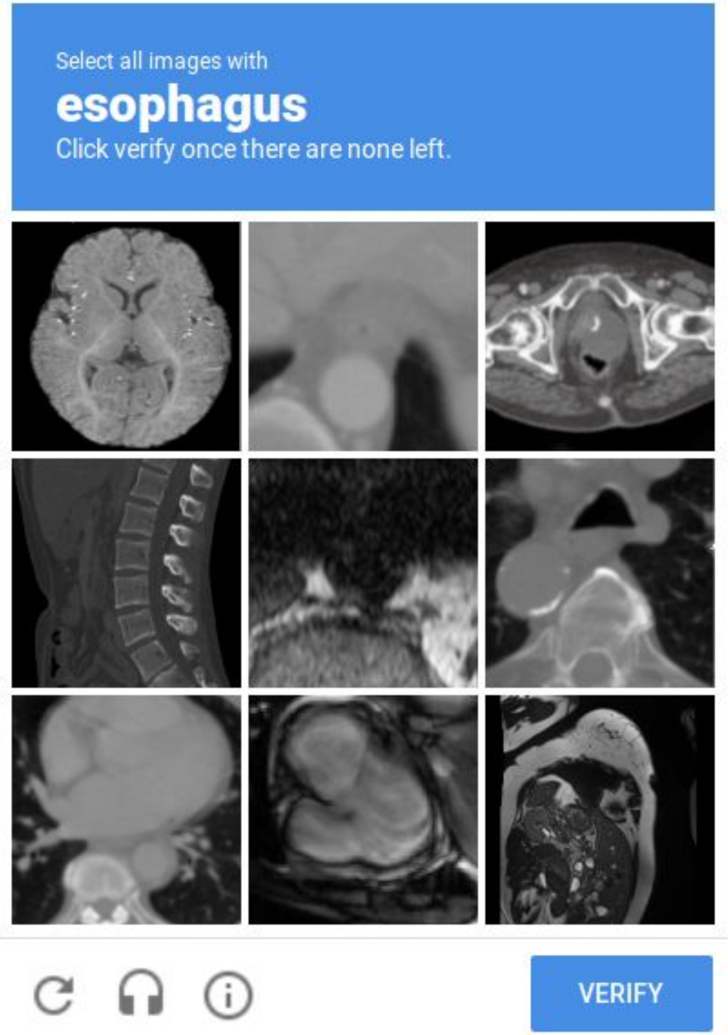
[Marin et al., CVPR 2019], [Tang et al., ECCV 2018],
[Lin et al., CVPR 2016], [Khoreva et al. CVPR 2017],
[Vernaza et al., CVPR 2017], [Kolesnikov and Lampert, ECCV 2016]
[Dai et al., CVPR 2015], [Bearman et al., ECCV 2016]
[Pathak et al., ICCV 2015], [Papandreou et al., ICCV 2015]

[Rajchl et al., TMI 2017]
[Bai et al., MICCAI 2017]
[Kervadec et al., MedIA]

Full annotations are much more problematic in medical imaging

Not anywhere close to the 10k images of Pascal VOC and the 5k of Cityscapes

Crowdsourcing?

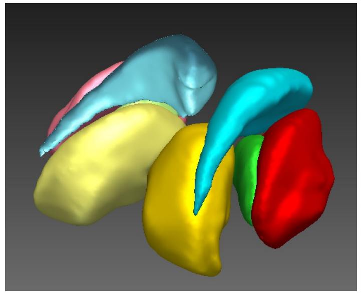


Full annotations are much more problematic in medical imaging

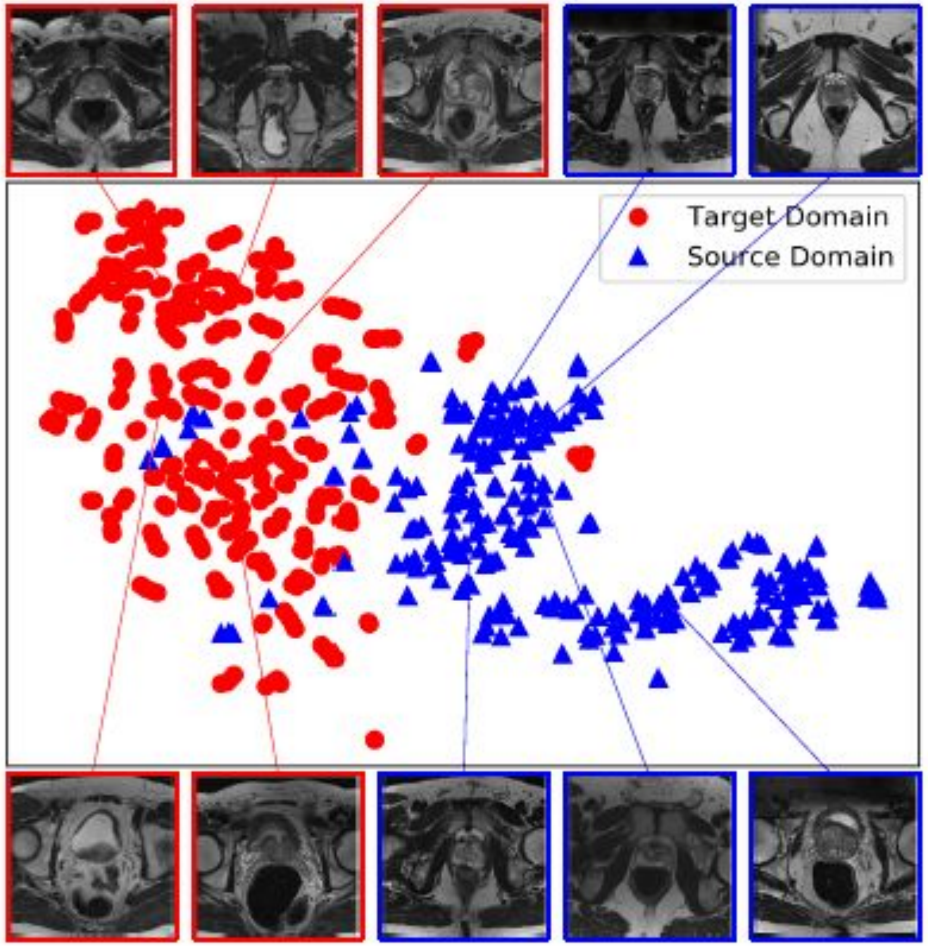
Not anywhere close to the 10k images of Pascal VOC and the 5k of Cityskapes

Crowdsourcing?

Dense 3D annotations: several hours (of radiologist time)

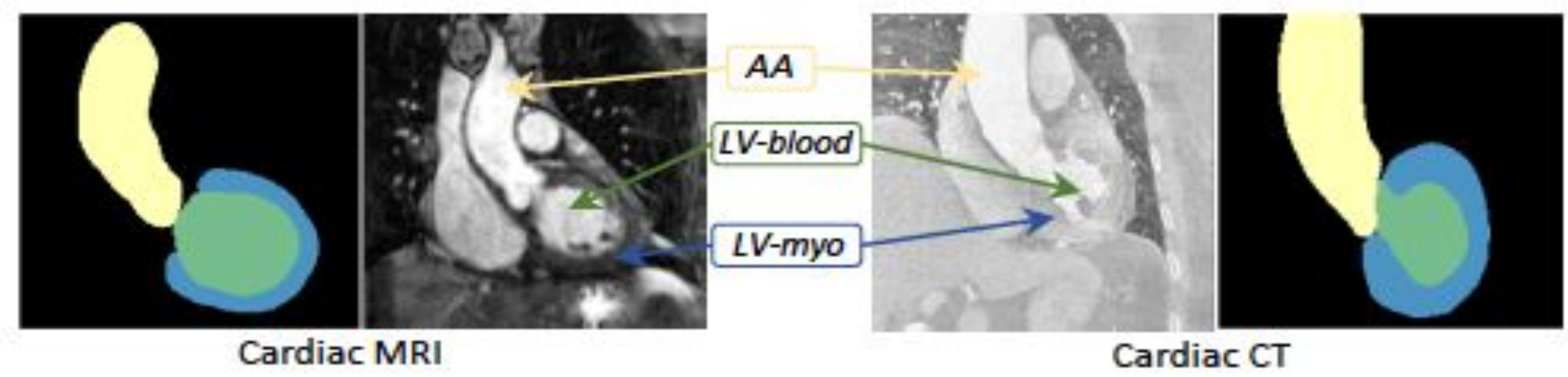


Domain shifts make things worse (even with full annotations in one domain)



[MRI Prostate segmentation: Figure from Zhu et al., Boundary-weighted Domain Adaptive Neural Network for Prostate MR Image Segmentation ArXiv 2019]

Domain shifts: within and across modalities

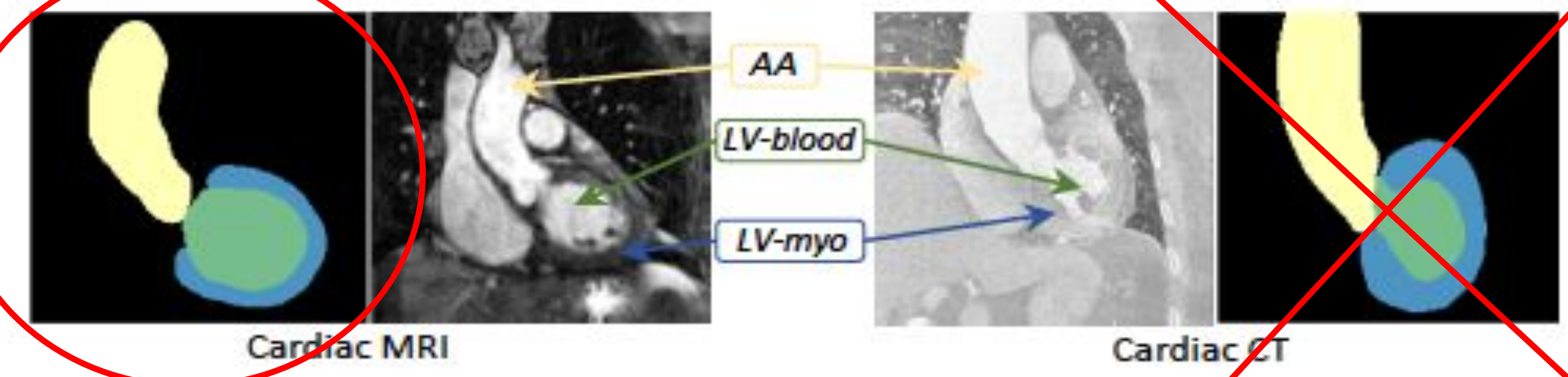


[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

Unsupervised domain adaptation

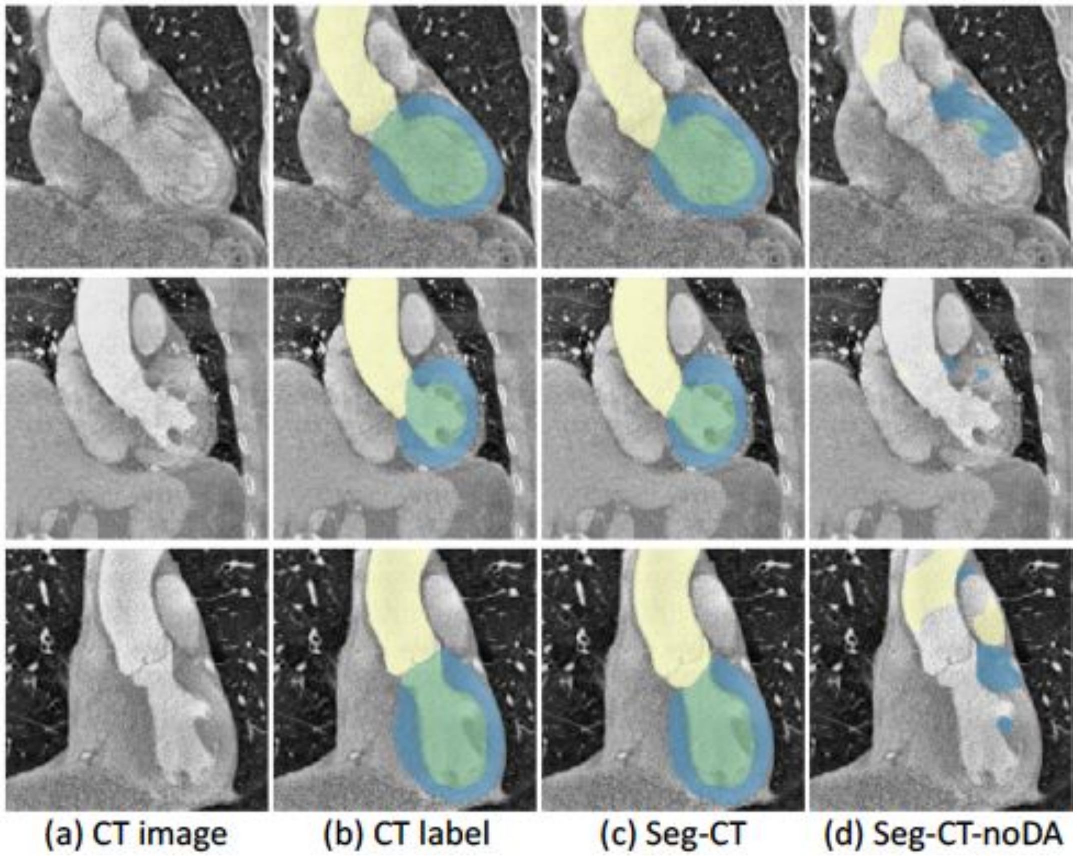
We have labels for the source domain

No labels for the target



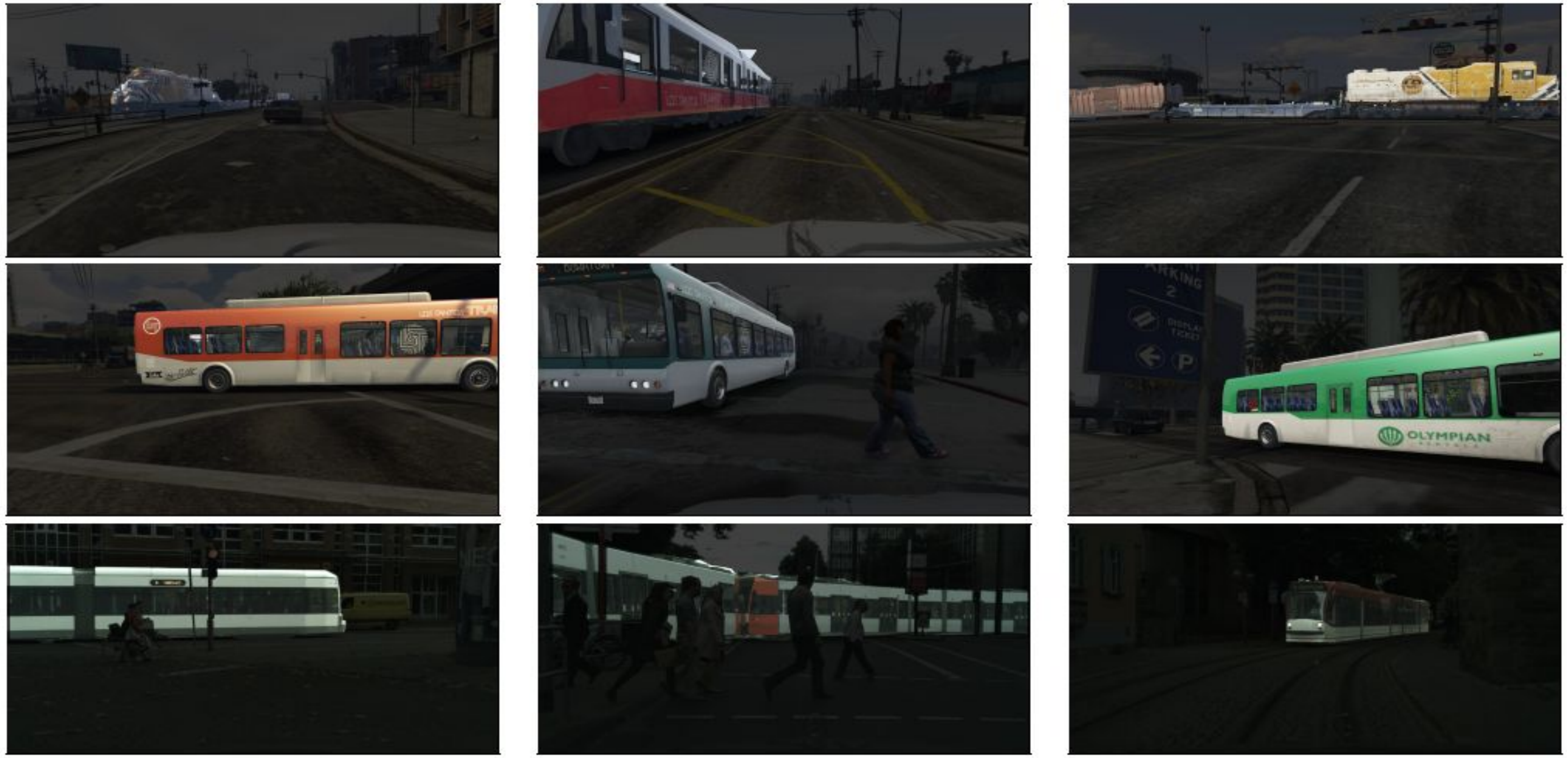
[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

Bad generalization to the target



[Images from Dou et al., PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation ArXiv 2018]

A lot of interest in vision as well:
Domain shifts are *everywhere* **BUT** we cannot label *everywhere*



"train"
GTA

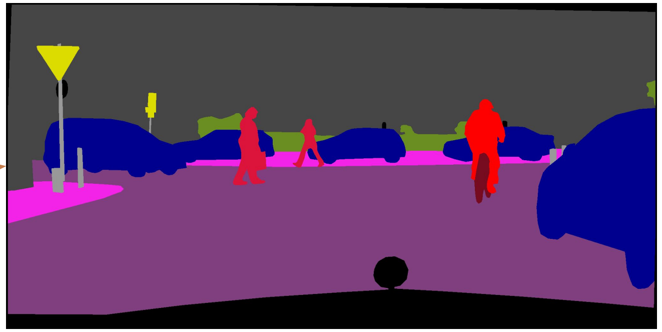
"bus"
GTA

"train"
Cityscapes

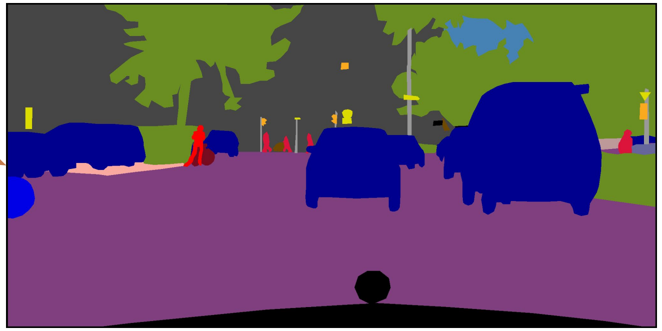
Figures from [Zhang et al., A Curriculum Domain Adaptation Approach to the Semantic Segmentation of Urban Scenes TPAMI 2019]

A lot of interest in vision as well:
Domain shifts are *everywhere* **BUT** we cannot label *everywhere*

road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain
sky	person	rider	car	truck	bus	train	motorcycle	bicycle	unlabeled



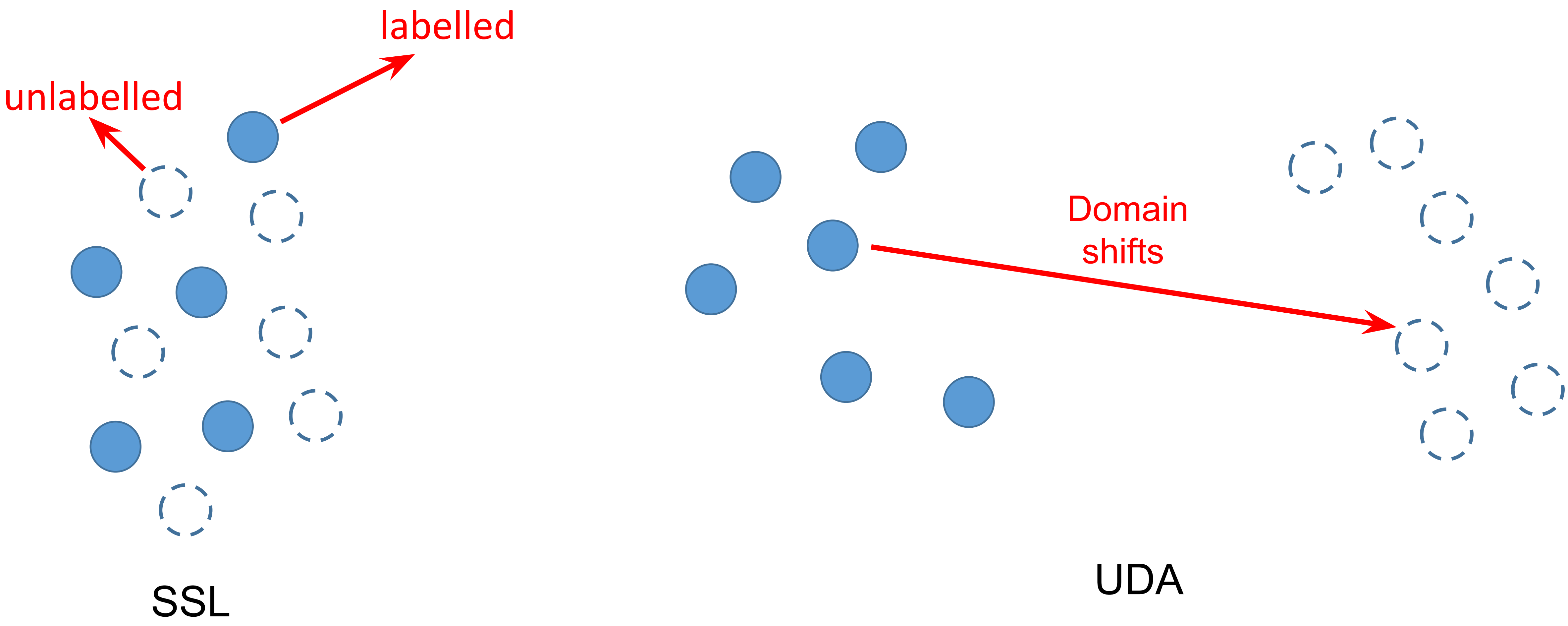
Frankfurt



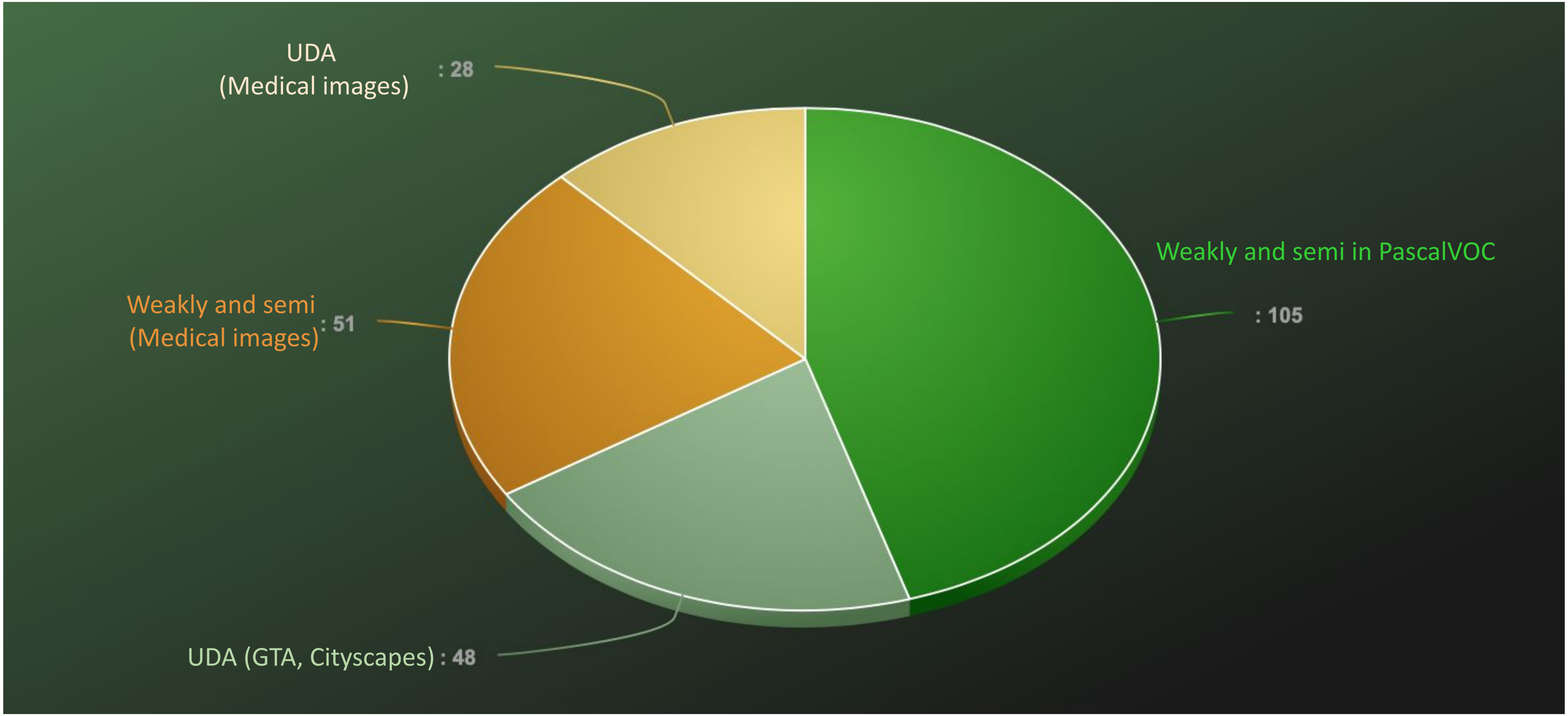
Zurich

Cityscapes (5000 images): labeling of 1 image takes 90 min at average [Cordt et al., CVPR 2016]

UDA = SSL + domain shift



Surprisingly in medical image analysis, we are behind



Semi/weak supervision in a nutshell:
We are leveraging **unlabelled** data with **priors**

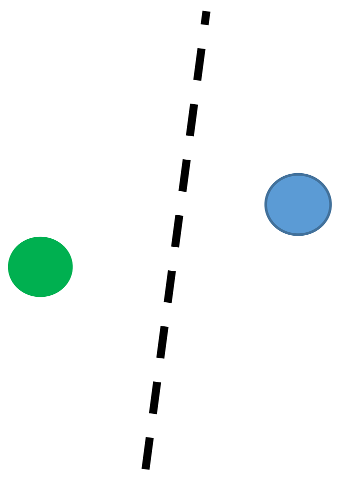
- Structure-driven priors: *Regularization (Part 1)*
- Knowledge- and data-driven priors (Part 2)

Part 1

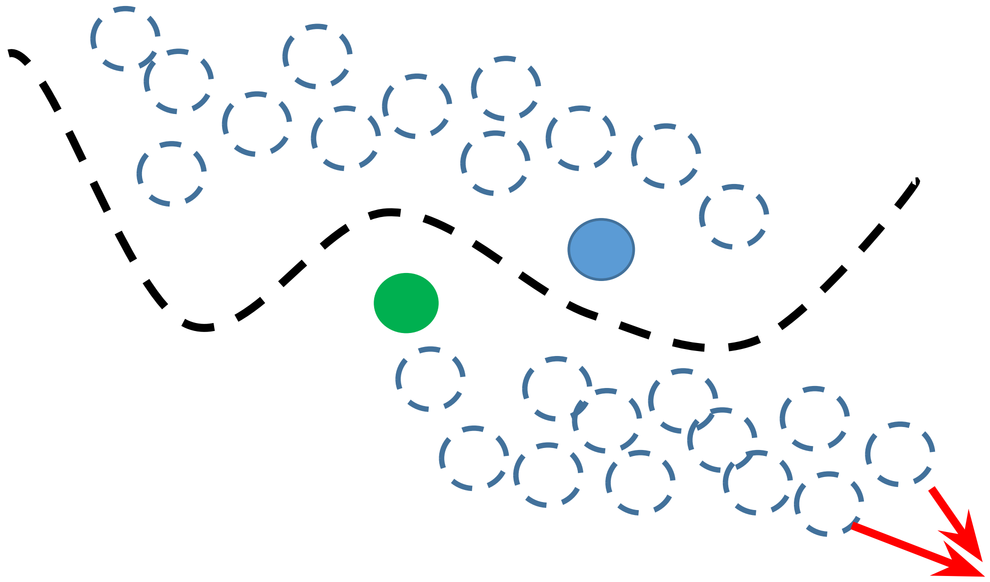
Regularization

Laplacian (and CRFs)

Semi-supervised learning (general form)

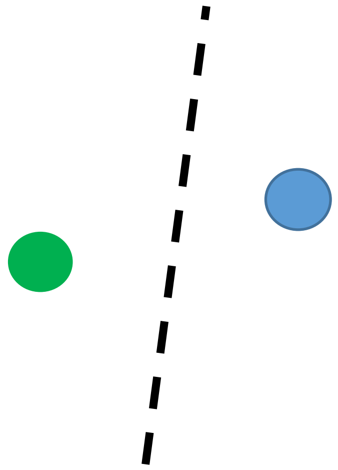


Only labeled points

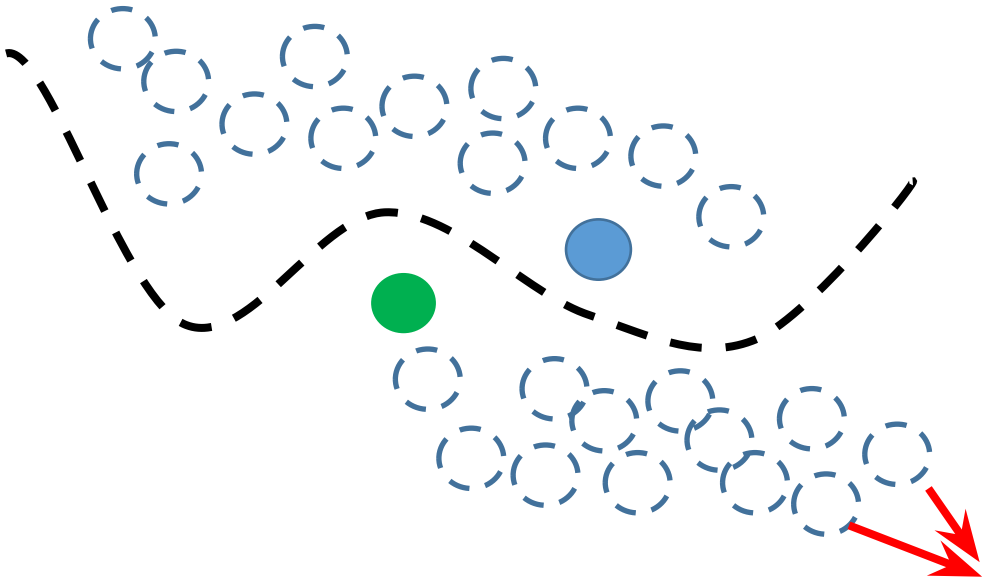


Unlabeled points

Semi-supervised learning (general form)



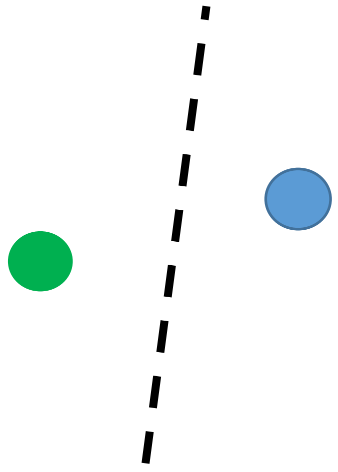
Only labeled points

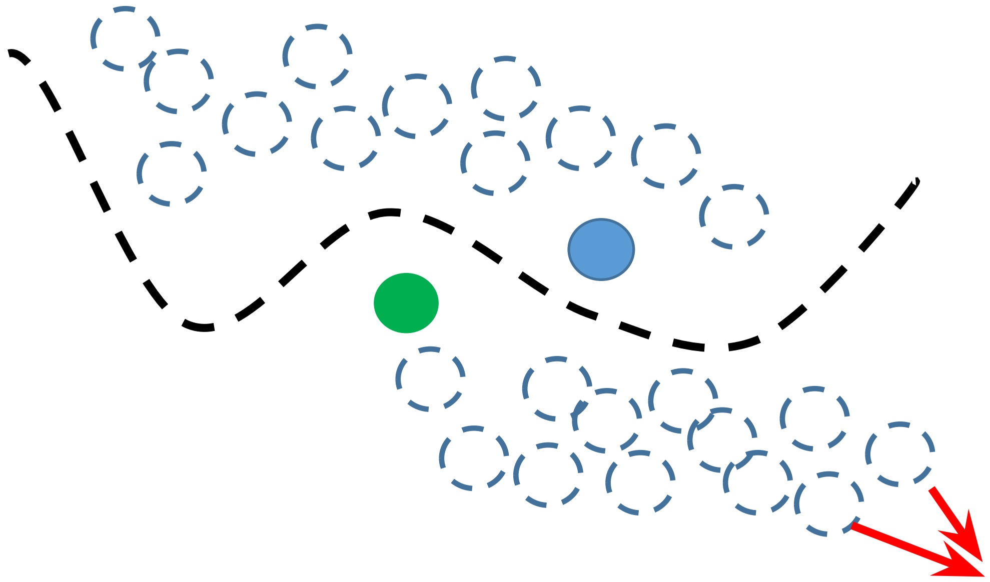


Unlabeled points

$$\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$$

Semi-supervised learning (general form)





$$\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$$

Labeled points

Unlabeled points

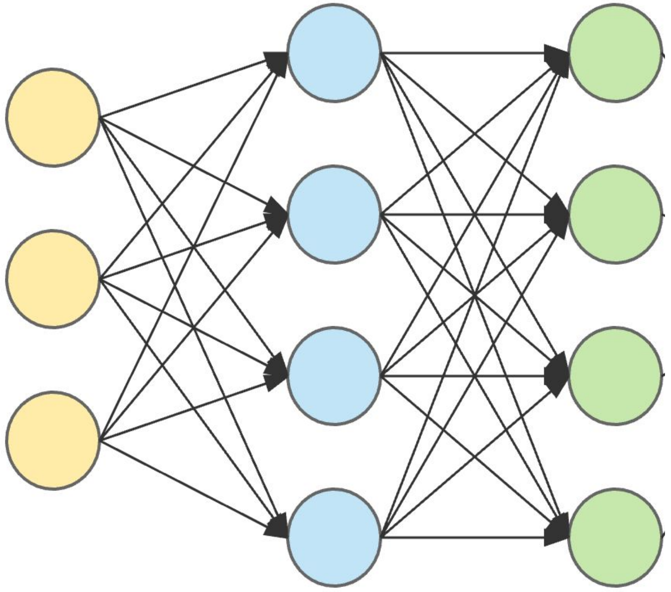
Semi-supervised learning (general form)

$$\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$$

Labels
(binary simplex vectors)

e.g.: cross-entropy

e.g.: simplex probability vectors
(*softmax outputs of the network*)



$$\left.\begin{array}{c} f_\theta^{p,1} \\ f_\theta^{p,2} \\ \vdots \\ f_\theta^{p,C} \end{array}\right\} \mathbf{f}_\theta^p = \mathbf{s}_\theta^p \in [0,1]^K$$

Semi-supervised learning (general form)

$$\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$$

Labels
(binary simplex vectors)

e.g.: **Laplacian**

$$w_{p,q} \|\mathbf{f}_\theta^p - \mathbf{f}_\theta^q\|^2$$

e.g.: cross-entropy

e.g.: simplex probability vectors
(*softmax outputs of the network*)

Semi-supervised learning (general form)

$$\sum_{p \in \mathcal{L}} l(\mathbf{f}_\theta^p, \mathbf{y}^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} g(\mathbf{f}_\theta^p, \mathbf{f}_\theta^q, w_{p,q})$$

Labels
(binary simplex vectors)

e.g.: cross-entropy

e.g.: **Laplacian**

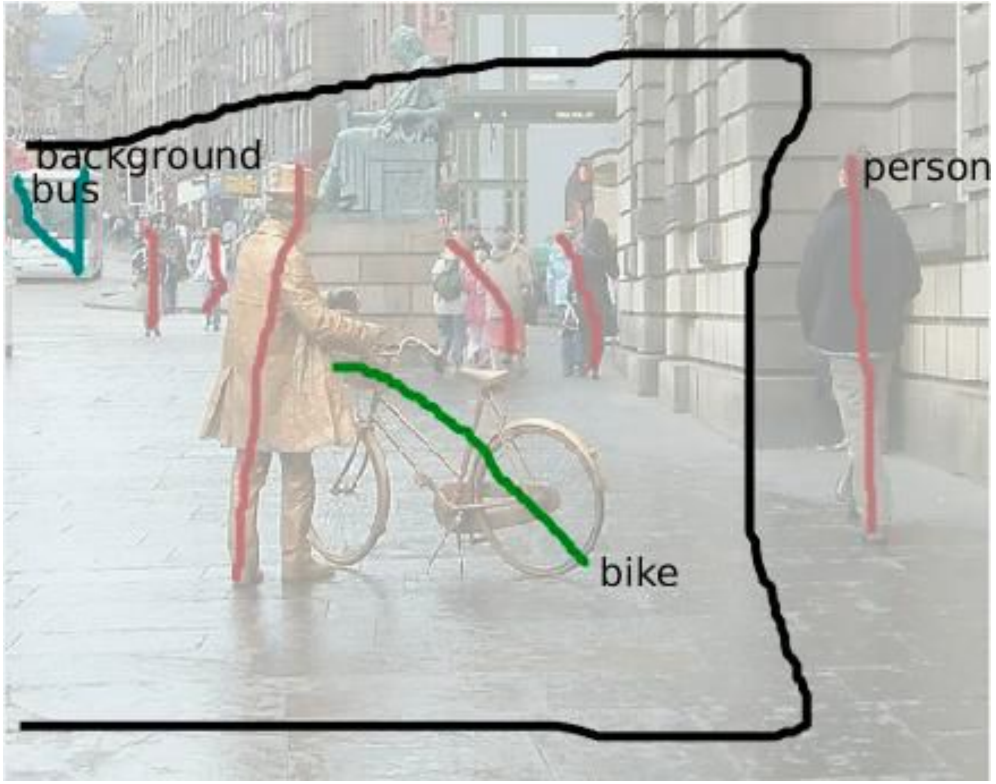
$$w_{p,q} \| \mathbf{f}_\theta^p - \mathbf{f}_\theta^q \|^2$$

e.g.: simplex probability vectors
(*softmax outputs of the network*)

- [Weston et al., Deep Learning via semi-supervised embedding, ICML 2008]
- [Belkin et al., Manifold regularization: a geometric framework for learning from Labeled and Unlabeled Examples, JMLR 2006]
- [Zhu et al., Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions, ICML 2003]

Semi-supervision loss for segmentation

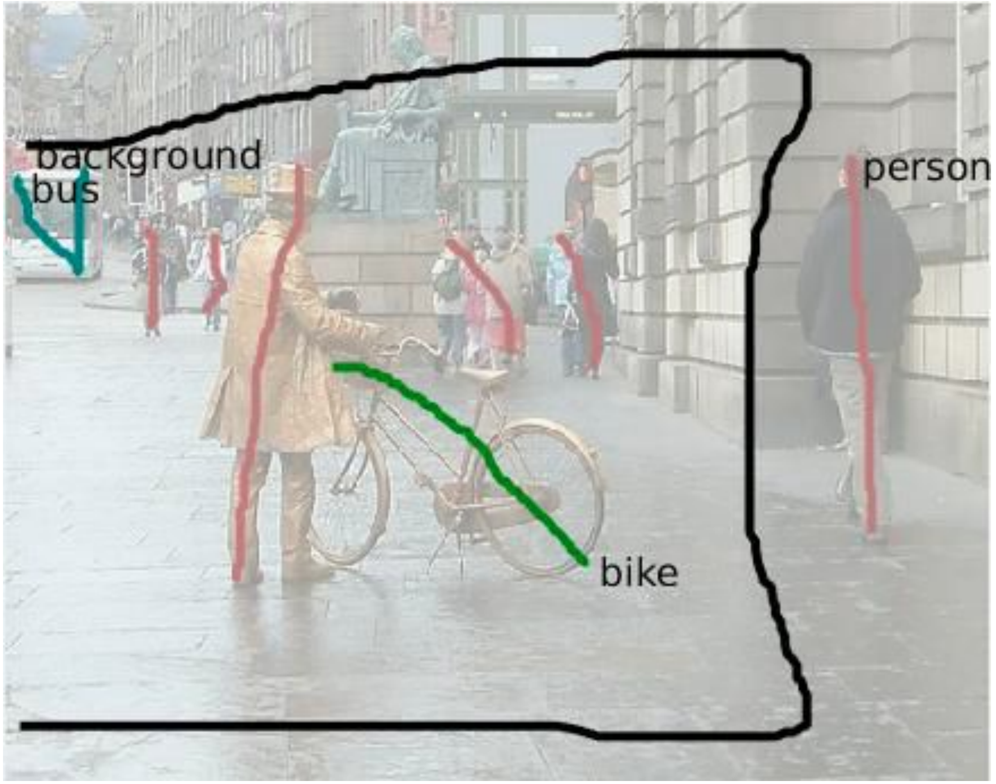
$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$



[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]

Semi-supervision loss for segmentation

$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$



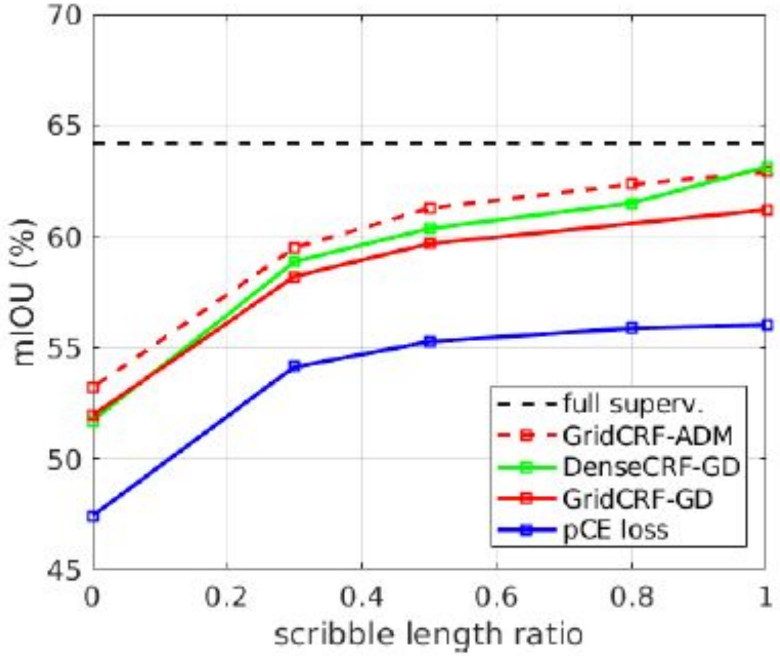
On the vertices of the simplex (binary variables), this is exactly the Potts model in Conditional Random Fields
(e.g., Dense CRFs)!

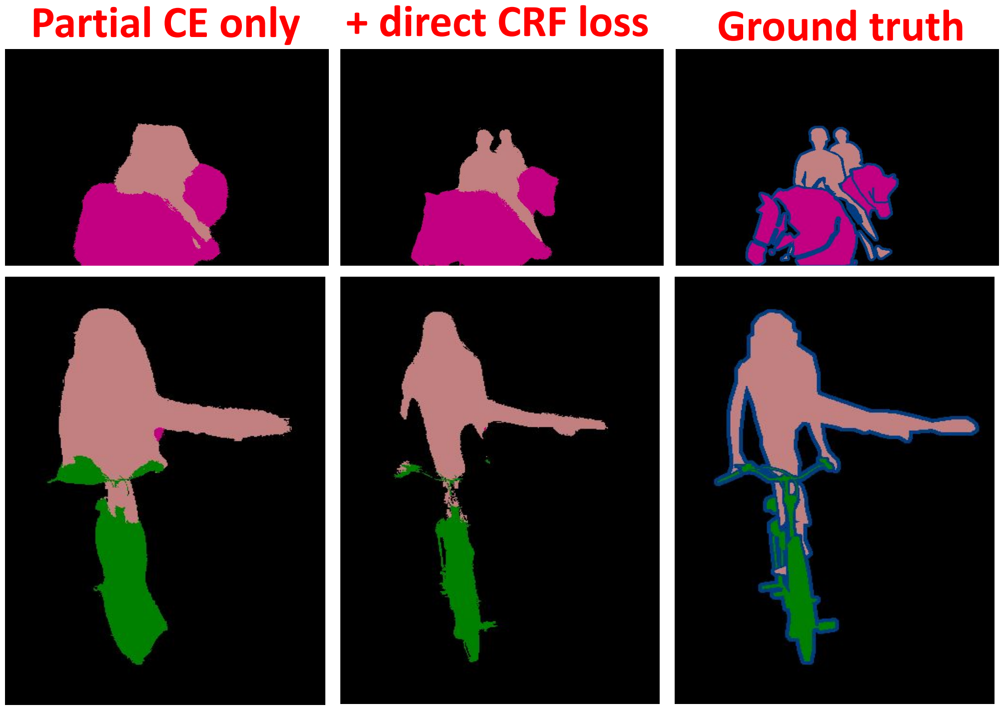
[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]

Semi-supervision loss for segmentation

$$\min_{\theta} \sum_{p\in\mathcal{L}} l(\mathbf{y}^p, \mathbf{s}^p_\theta) + \sum_{p,q\in\mathcal{L}\cup\mathcal{U}} w_{p,q} \|\mathbf{s}^p_\theta - \mathbf{s}^q_\theta\|^2$$

SGD

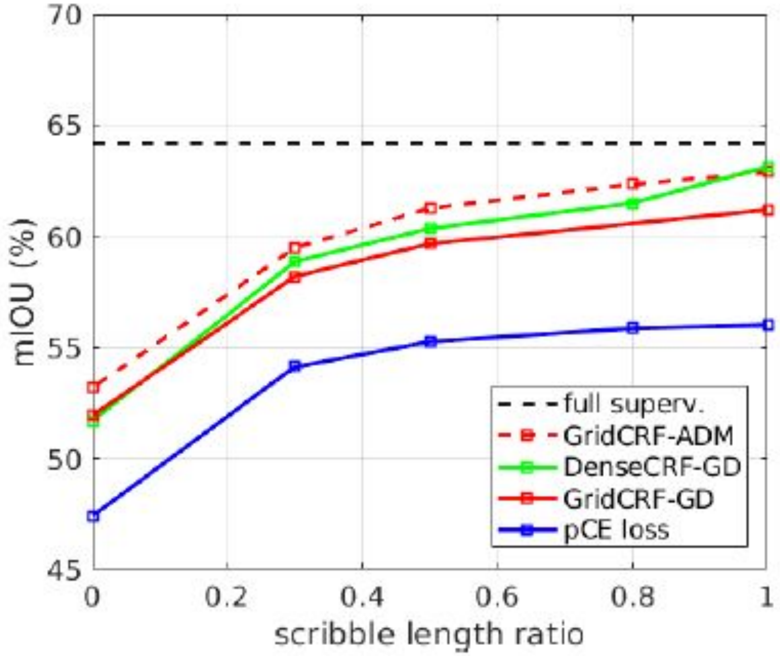




[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]
[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Semi-supervision loss for segmentation

$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}_\theta^p - \mathbf{s}_\theta^q\|^2$$



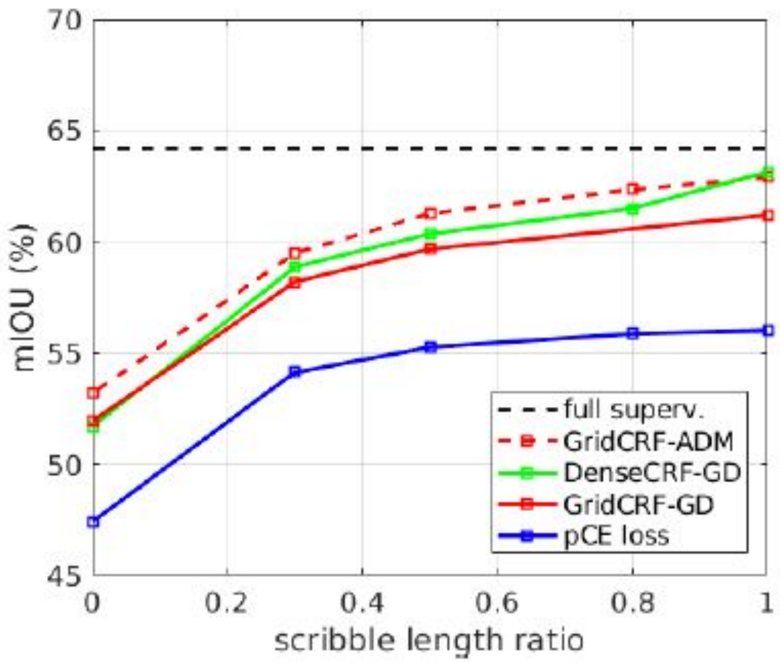
The exciting part in this plot:

Dense CRF with SGD gets you **97.6%** of full supervision performance with **3%** of the labels!

[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]
[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Semi-supervision loss for segmentation

$$\min_{\theta} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}^p_\theta) + \sum_{p,q \in \mathcal{L} \cup \mathcal{U}} w_{p,q} \|\mathbf{s}^p_\theta - \mathbf{s}^q_\theta\|^2$$



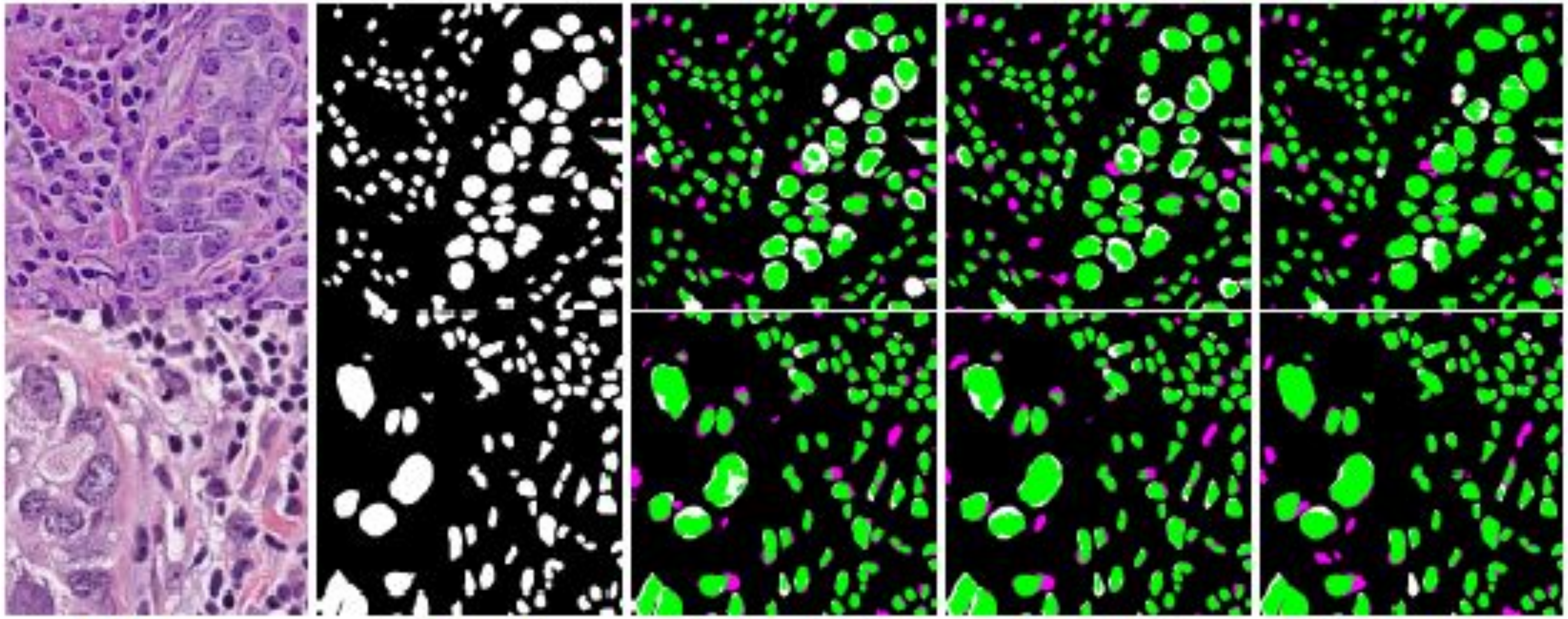
The disturbing part (for those who know classical CRFs):

Dense CRF is not supposed to be better than grid CRF

[Tang et al., On regularized losses for weakly supervised segmentation, ECCV 2018]
[Marin et al., Beyond gradient descent for regularized segmentation losses, CVPR 2019]

Some applications of CRF loss in MICCAI

White (FN); Magenta (FP); Green (TP)



(a) images (b) gt masks (c) without CRF (d) with CRF (e) full annotation

- Figures from Qu et al., Weakly Supervised Deep Nuclei Segmentation using Points Annotation in Histopathology Images, MIDL 2019 [Histology, point annotation]

- Ji et al., Scribble-Based Hierarchical Weakly Supervised Learning for Brain Tumor Segmentation, MICCAI 2019 [Brain tumor images, scribble annotations]

Regularization

entropy

Entropy minimization for SSL

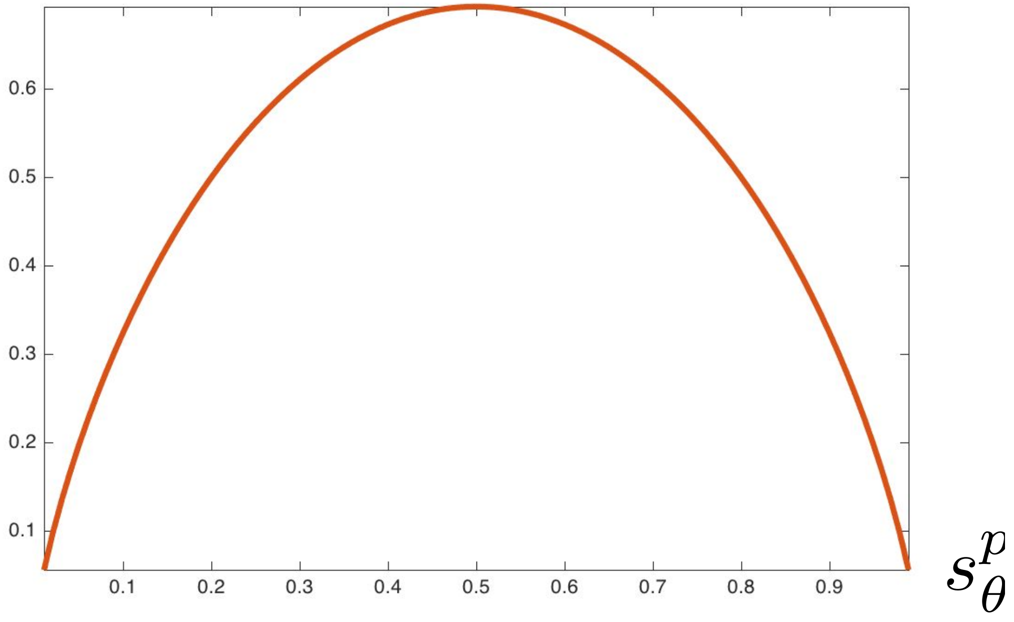
$$\min_{\theta} - \sum_{p \in \mathcal{L}} \sum_{c=1}^{C} y^{p,c} \log s_{\theta}^{p,c} - \sum_{p \in \mathcal{U}} \sum_{c=1}^{C} s_{\theta}^{p,c} \log s_{\theta}^{p,c}$$

Shannon Entropies: “unsupervised cross-entropies (with unknown labels)”

- Grandvalet & Bengio, Semi-supervised learning by entropy minimization, NIPS 2005
- Gomes et al., Discriminative clustering by regularized information maximization, NIPS 2010

Effect of the entropy (why is it good for SSL?):
It makes the predictions confident (like cross-entropy)

$$-s_\theta^p \log s_\theta^p - (1 - s_\theta^p)\log(1 - s_\theta^p)$$

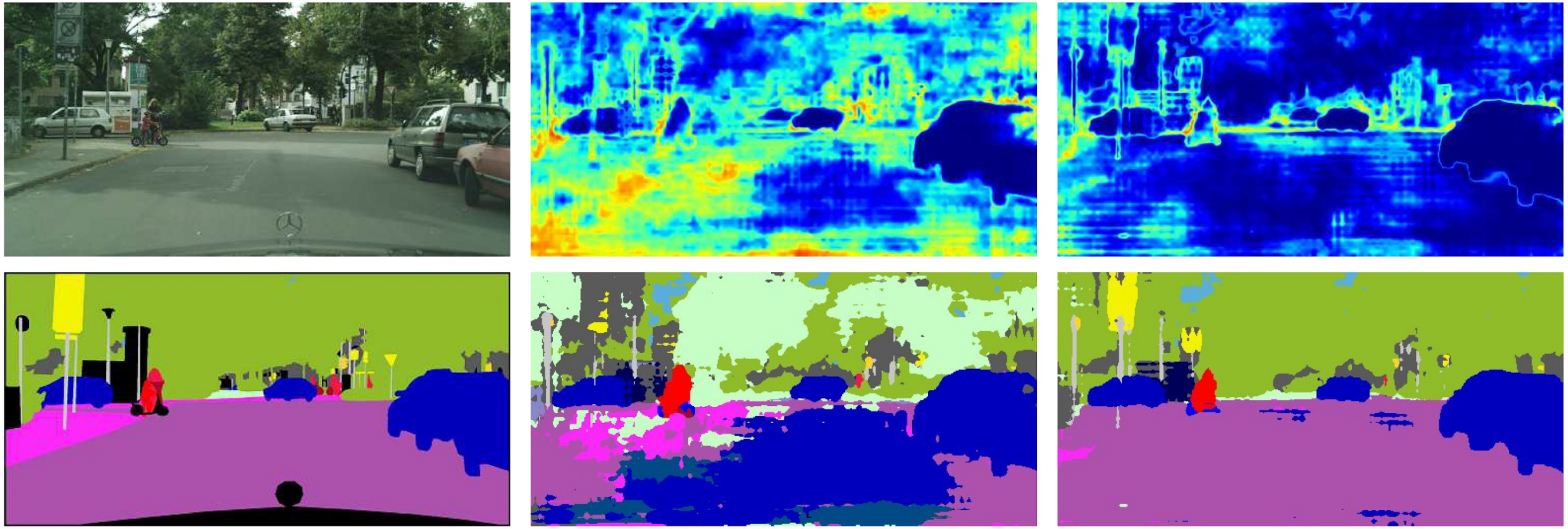


Entropy minimization for UDA

Input image + GT

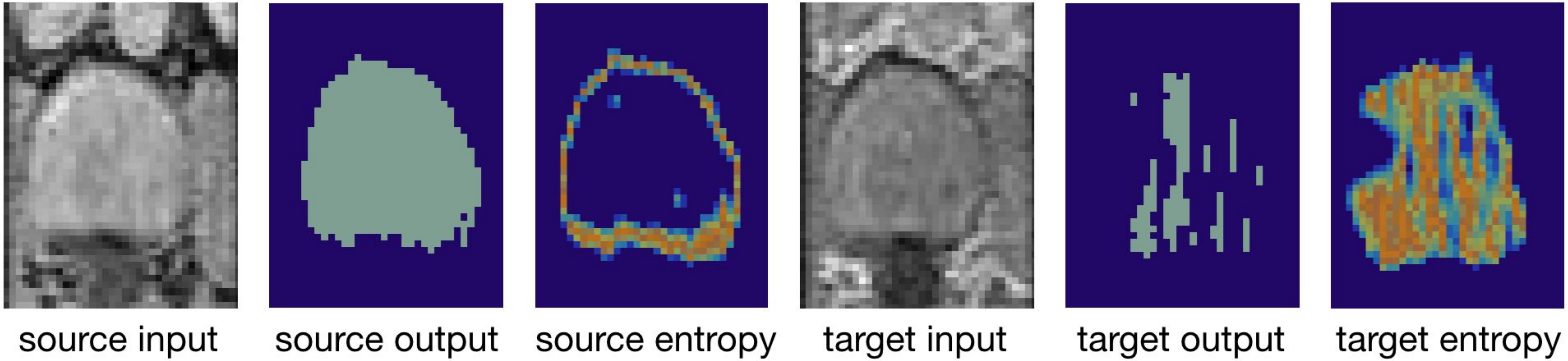
Without adaptation

Entropy minimization



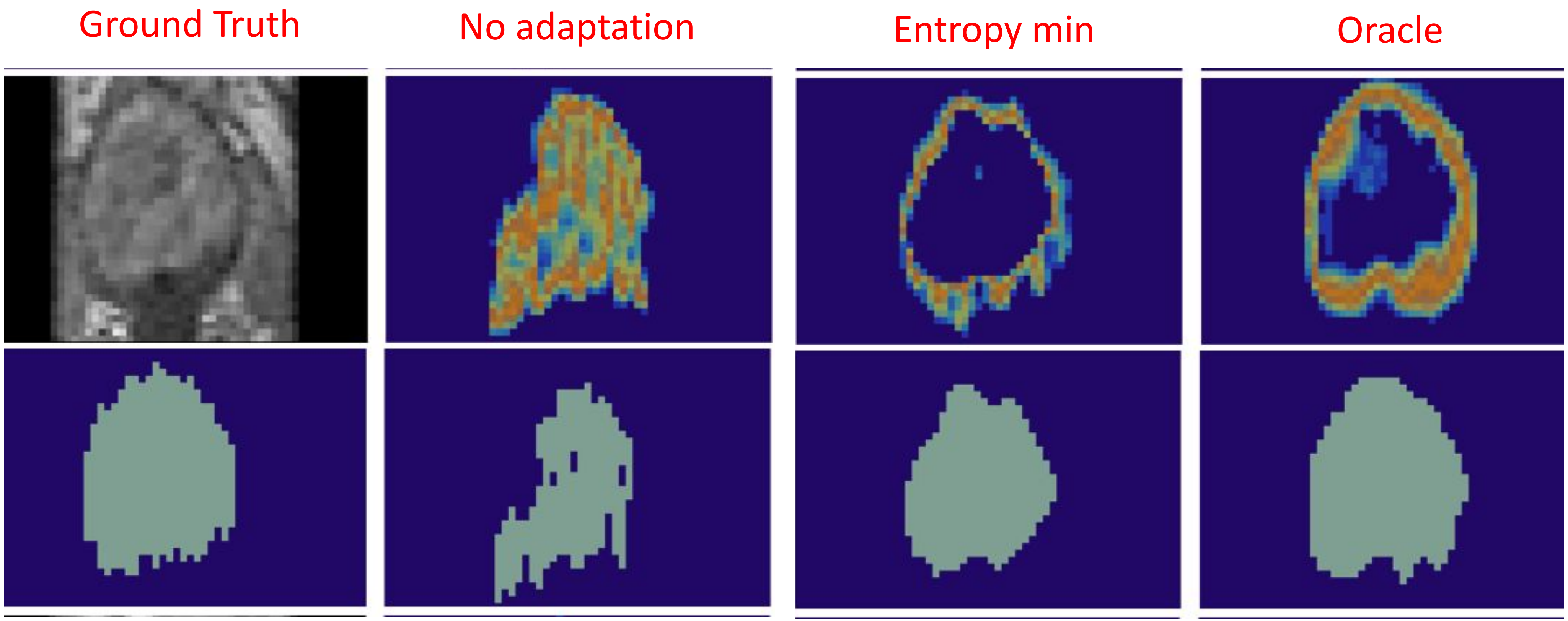
Images from Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation, CVPR 2019

Entropy minimization for UDA



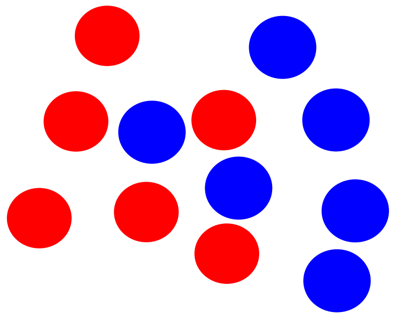
Images from Bateson et al., Source-relaxed domain adaptation for segmentation, MICCAI 2020

Entropy minimization for UDA



Images from Bateson et al., Source-relaxed domain adaptation for segmentation, MICCAI 2020

Why entropy minimization is good
(It increases the margin between the classes)



High entropy
(low confidence)

Low entropy
(high confidence)

Effect of the entropy (why is it good for SSL?): It increases the margin between the classes

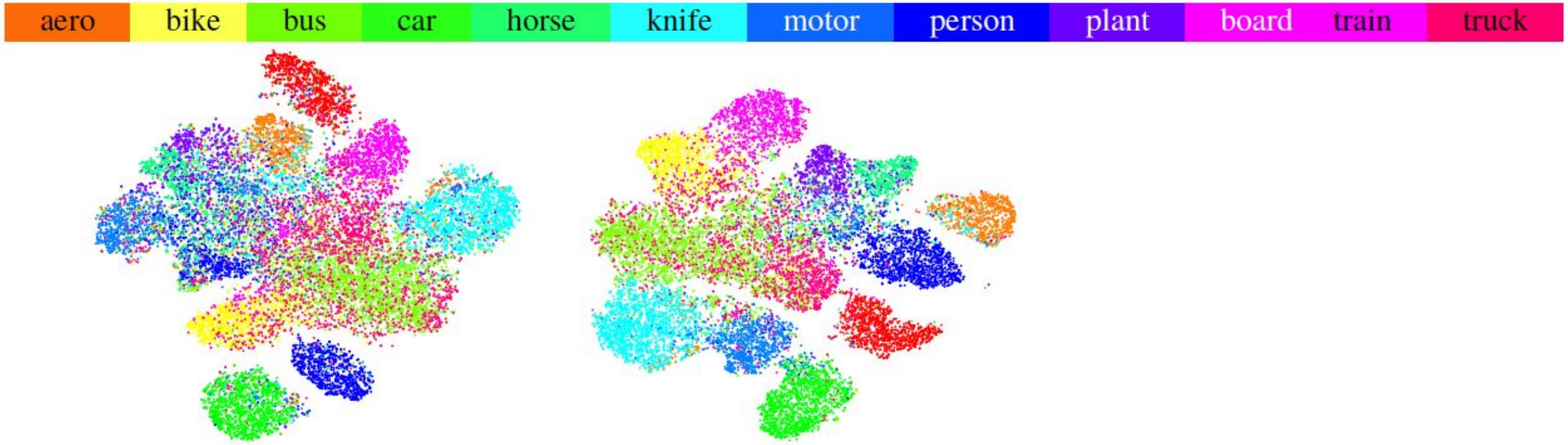
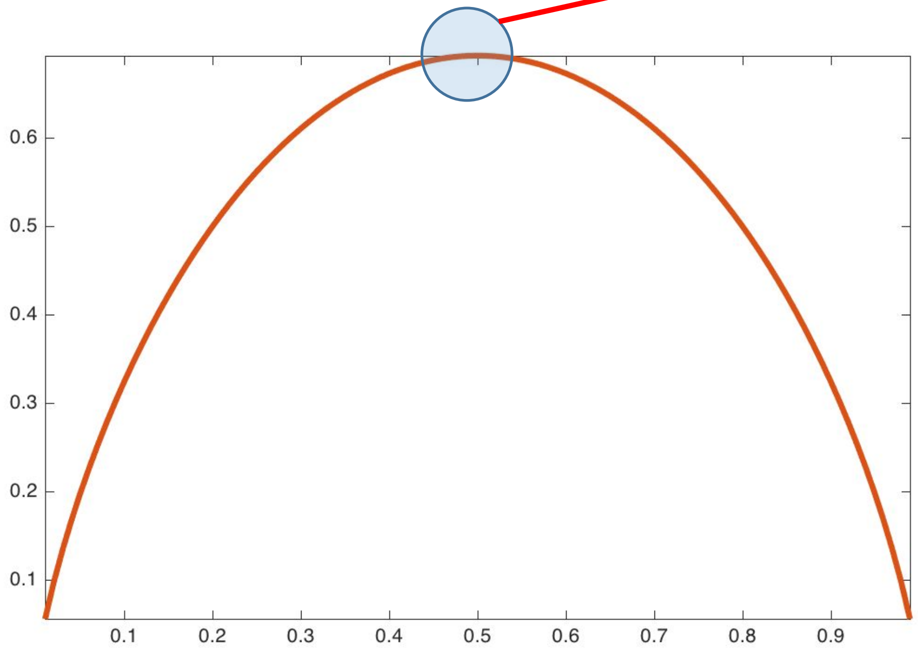


Image classification UDA on VisDA17 data set: Feature visualization for source model (left) and *min-entropy (lower bound on Shannon)* minimization (right) - equivalent to self training (clarified in the next slide)

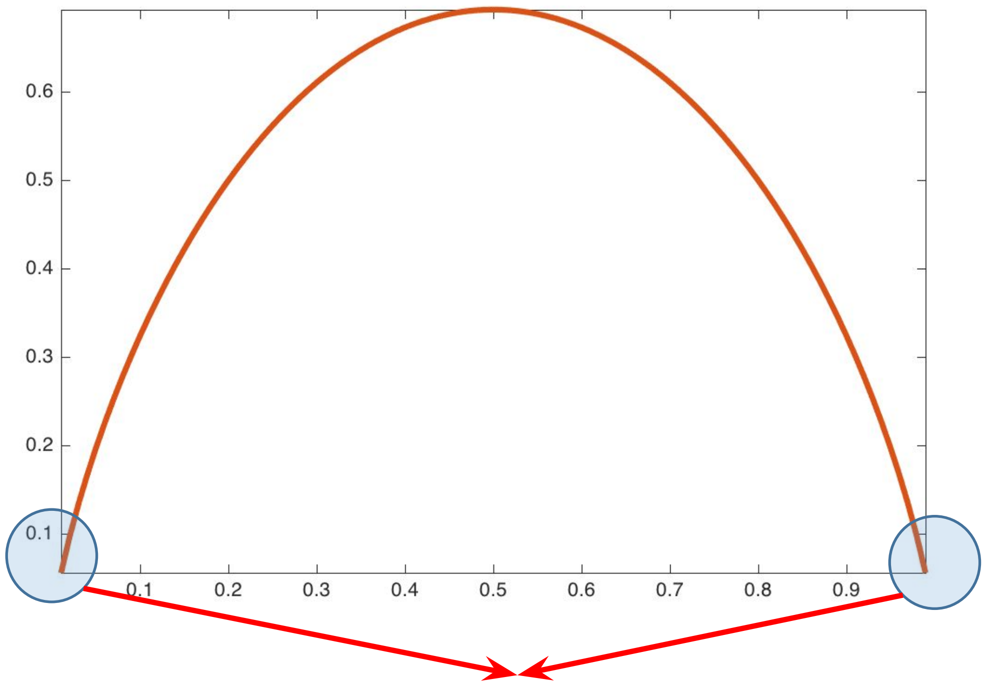
Figures from Zou et al., Confidence regularized self training, ICCV 2019

Difficulty of optimizing entropy

how is gradient here?



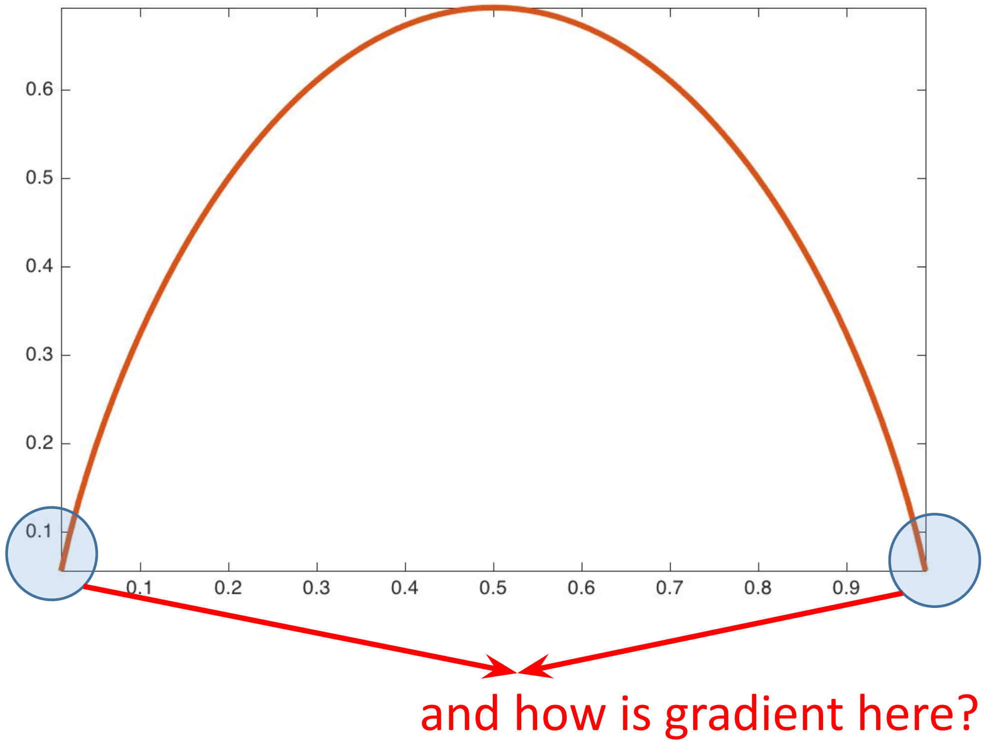
Difficulty of optimizing entropy



and how is gradient here?

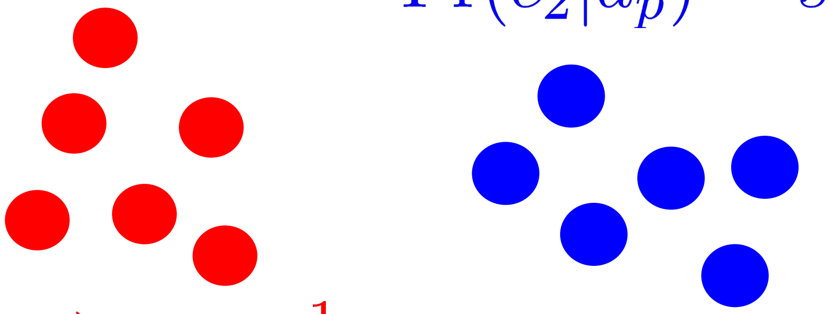
Difficulty of optimizing entropy

Typically we add other cues to facilitate optimization and avoid trivial solutions (more on this later)



Avoiding the trivial solutions of entropy minimization

$$\Pr(c_2|d_p) = s^{p,2} = 1 - s^{p,1} = 1$$



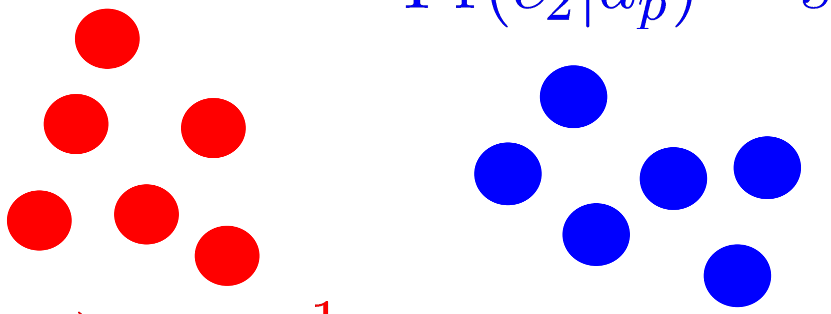
$$\Pr(c_1|d_p) = s^{p,1} = 1$$

Softmax output

Note: Ignoring the network parameters to simplify notation

Avoiding the trivial solutions of entropy minimization

$$\Pr(c_2|d_p) = s^{p,2} = 1 - s^{p,1} = 1$$

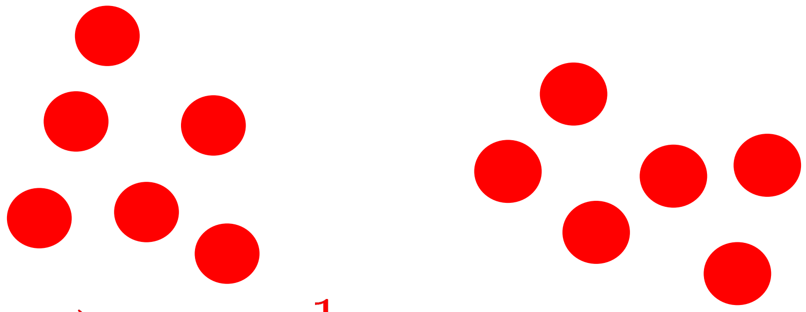


$$\Pr(c_1|d_p) = s^{p,1} = 1$$

Min entropy

(max confidence)

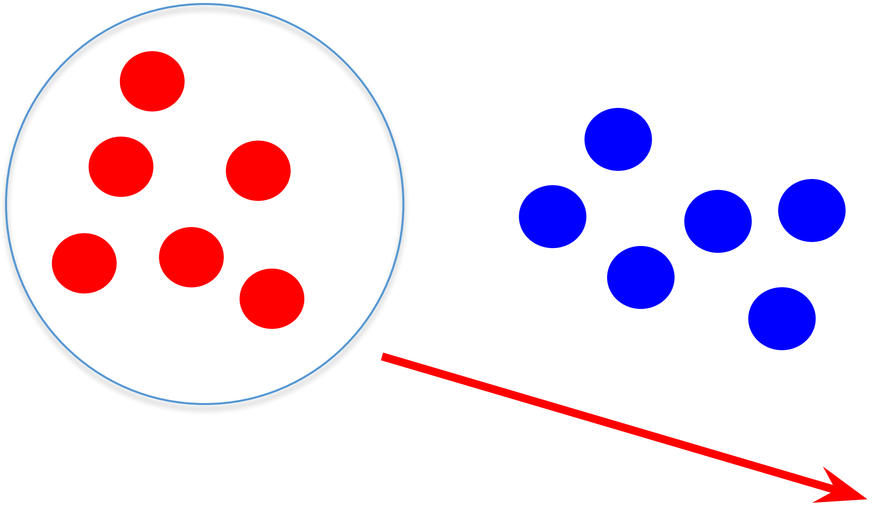
Avoiding the trivial solutions of entropy minimization



$$\Pr(c_1|d_p) = s^{p,1} = 1$$

This bad solution also has a minimum entropy!!!

Avoiding the trivial solutions of entropy minimization



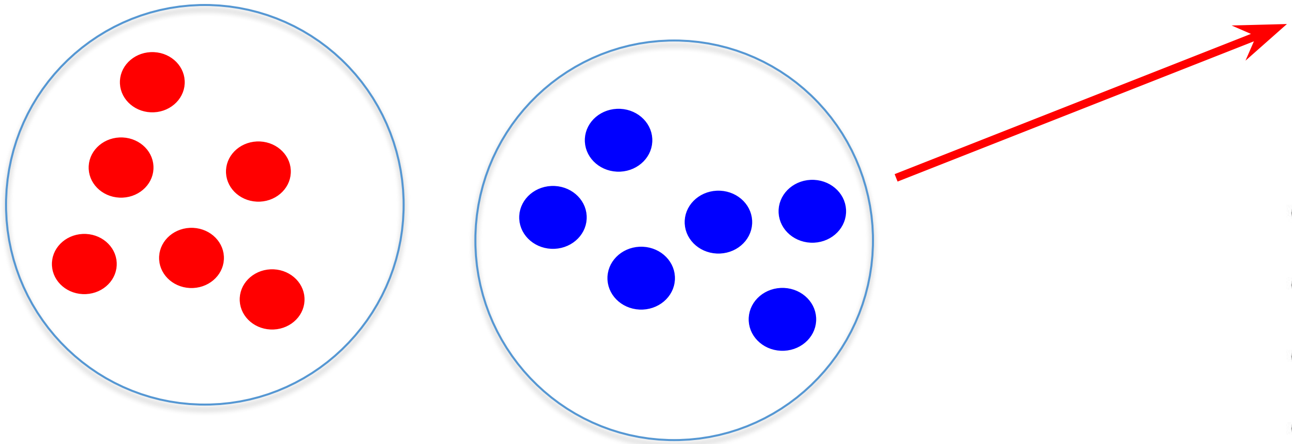
Marginal probabilities of the labels

-- Class proportion

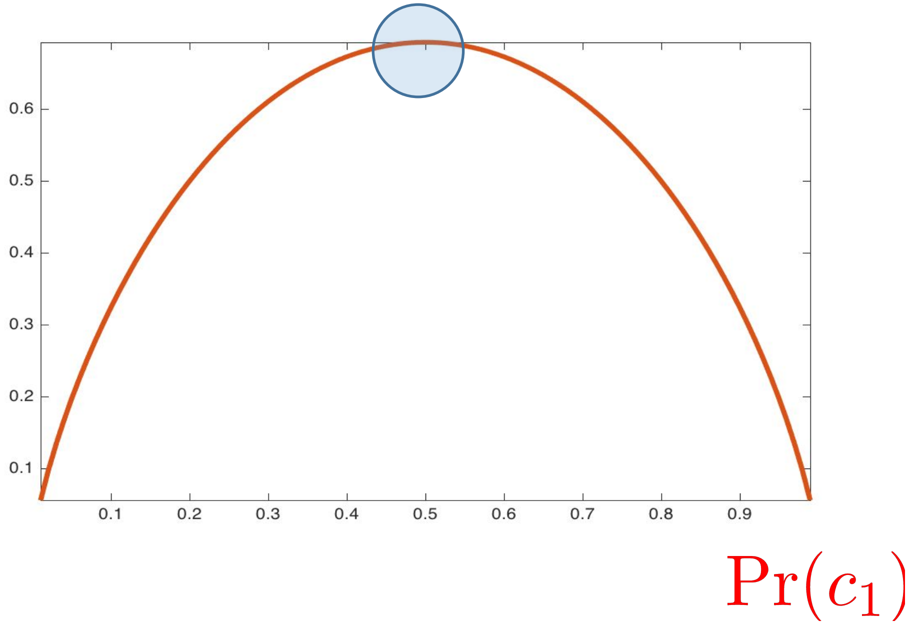
-- Region size (normalized) in segmentation

$$\Pr(c_1) \propto \sum_p s^{p,1}$$

Avoiding the trivial solutions of entropy minimization



Balanced solution **maximizes** the entropy of **label marginal**



Maximizing the mutual info (MI)
(between data points and their latent labels)

$I(X,Y) = H(Y) - H(Y/X)$

MI = ***Entropy*** (label marginal) – ***Entropy*** (posterior)

Standard and old in clustering, e.g.,:

Gomes et al., Discriminative clustering by regularized information maximization, NIPS 2010

Maximizing the mutual info (MI) (between data points and their latent labels)

MI = ***Entropy*** (label marginal) – ***Entropy*** (posteriors)

||

Up to a constant

KL (label marginal || uniform)

Standard and old in clustering:
Gomes et al., Discriminative clustering by regularized information maximization, NIPS 2010

Maximizing the mutual info (MI) (between data points and their latent labels)

MI = ***Entropy*** (label marginal) – ***Entropy*** (posteriors)

||

Up to a constant

KL (label marginal || uniform)

Other priors (learned, known)

Standard and old in clustering:

Gomes et al., Discriminative clustering by regularized information maximization, NIPS 2010

Maximizing the mutual info (MI)
(between data points and their latent labels)

MI = ***Entropy*** (label marginal) – ***Entropy*** (posteriors)

Up to a constant

KL (label marginal || uniform)

Other distances, relaxation of equality constraints

Standard and old in clustering:
Gomes et al., Discriminative clustering by regularized information maximization, NIPS 2010

Maximizing the mutual info (MI) (between data points and their latent labels)

Semi-supervised learning, e.g.

[Berthelot et al., NeurIPS'19]

[Kervadec et al., MedIA'19]

Few-shot learning, e.g.,

[Boudiaf et al., NeurIPS'20]

[Dhillion et al., ICLR'20]

Maximizing MI or its parts/proxies/generalizations is **SOTA almost everywhere!**

Unsupervised domain adaptation, e.g.,

Liang et al., ICML'20

Bateson et al., MICCAI'20

Deep clustering

&

Unsupervised Representation Learning, e.g.,

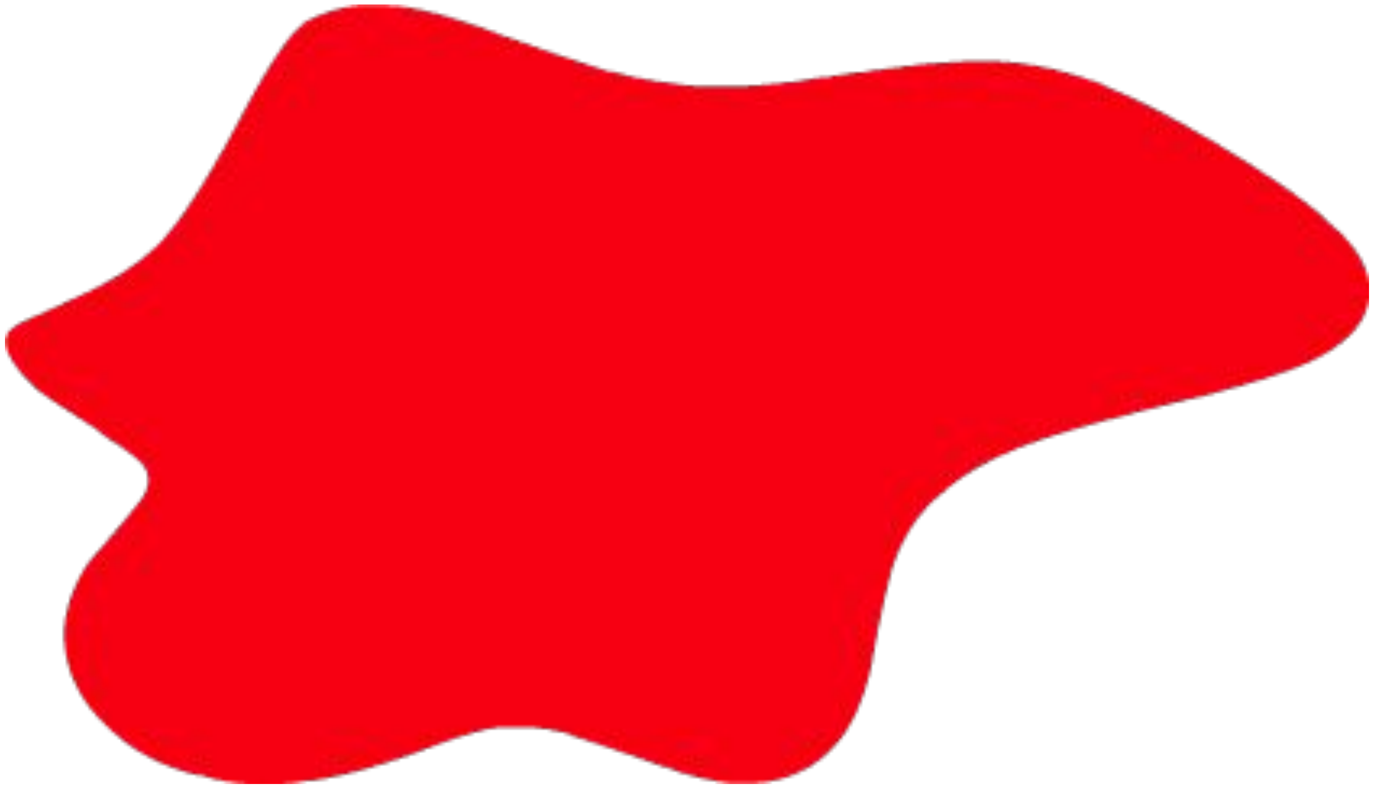
Asano et al., ICLR'20

Jabi et al., TPAMI'20

Constrained CNNs

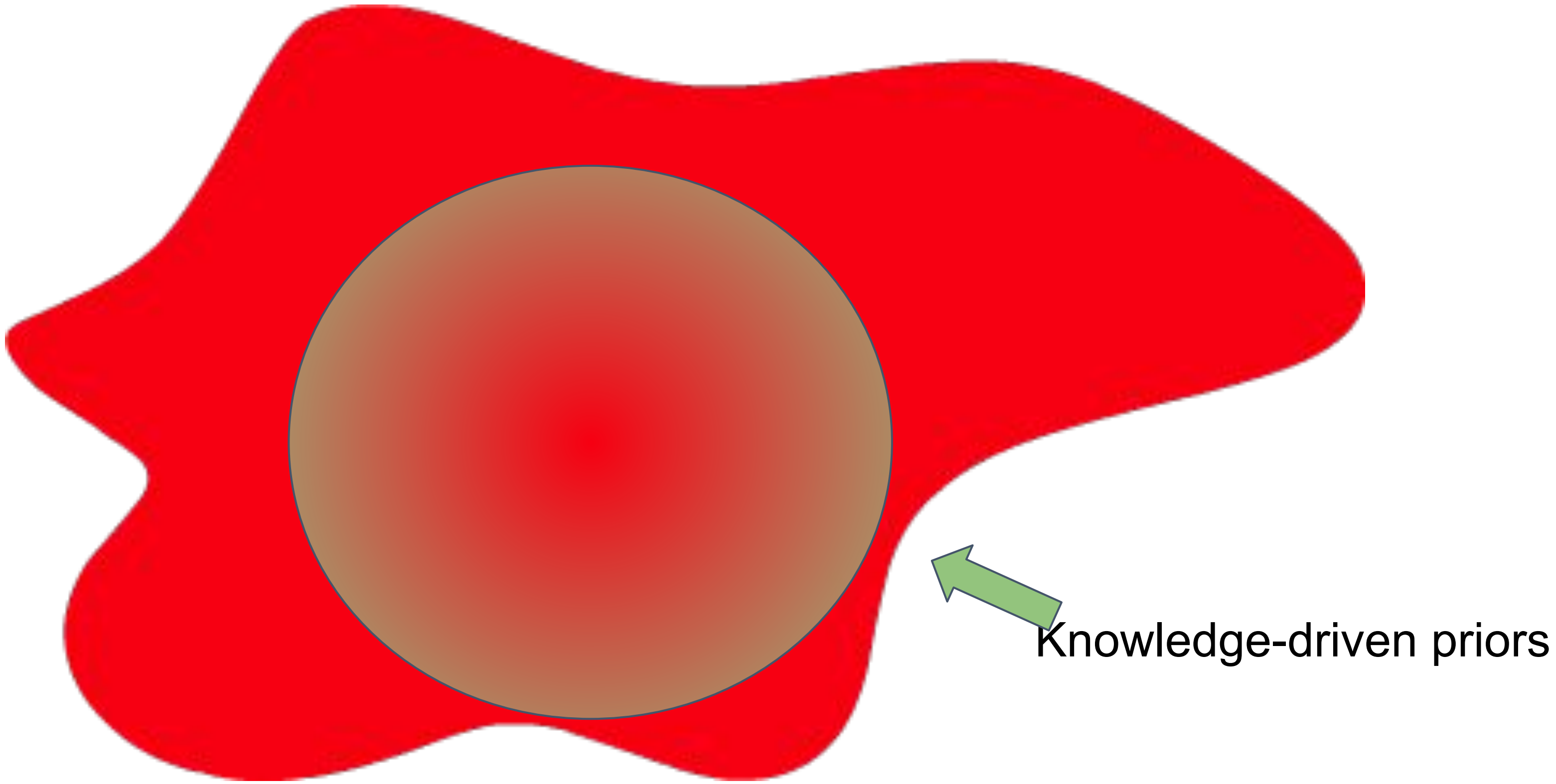
Constrained optimization (in CNNs)

Knowledge vs data driven priors



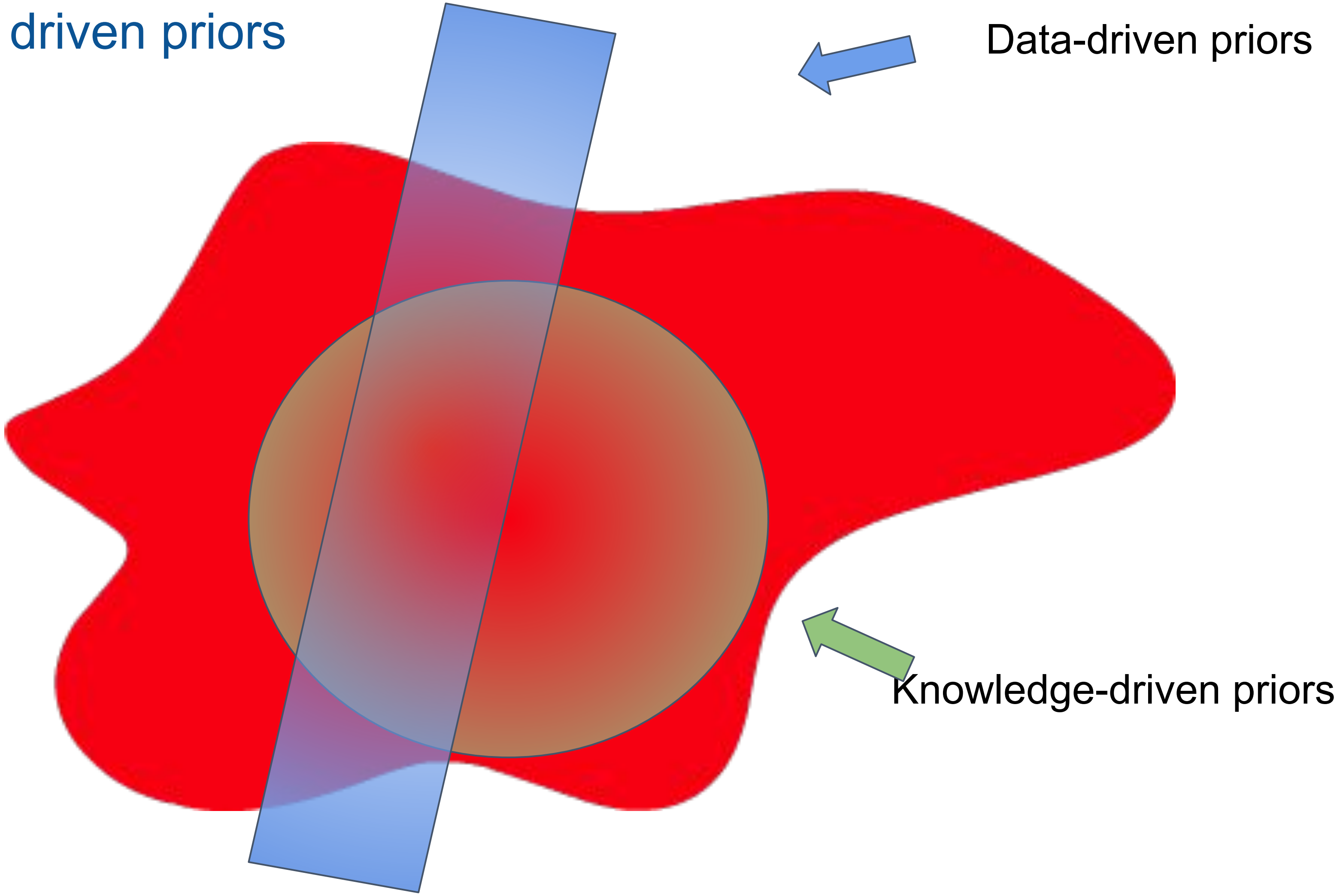
Constrained optimization (in CNNs)

Knowledge vs data driven priors



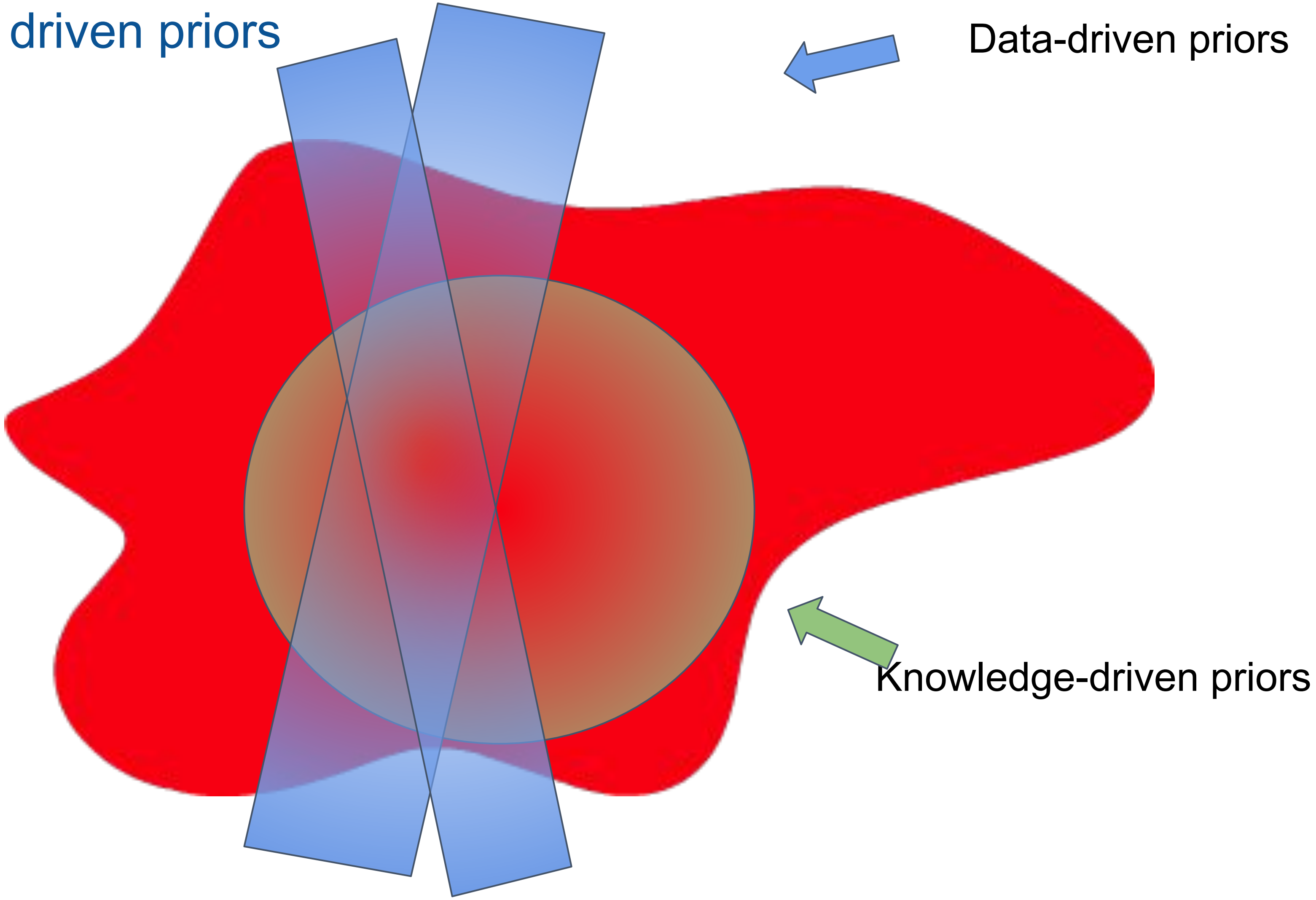
Constrained optimization (in CNNs)

Knowledge vs data driven priors



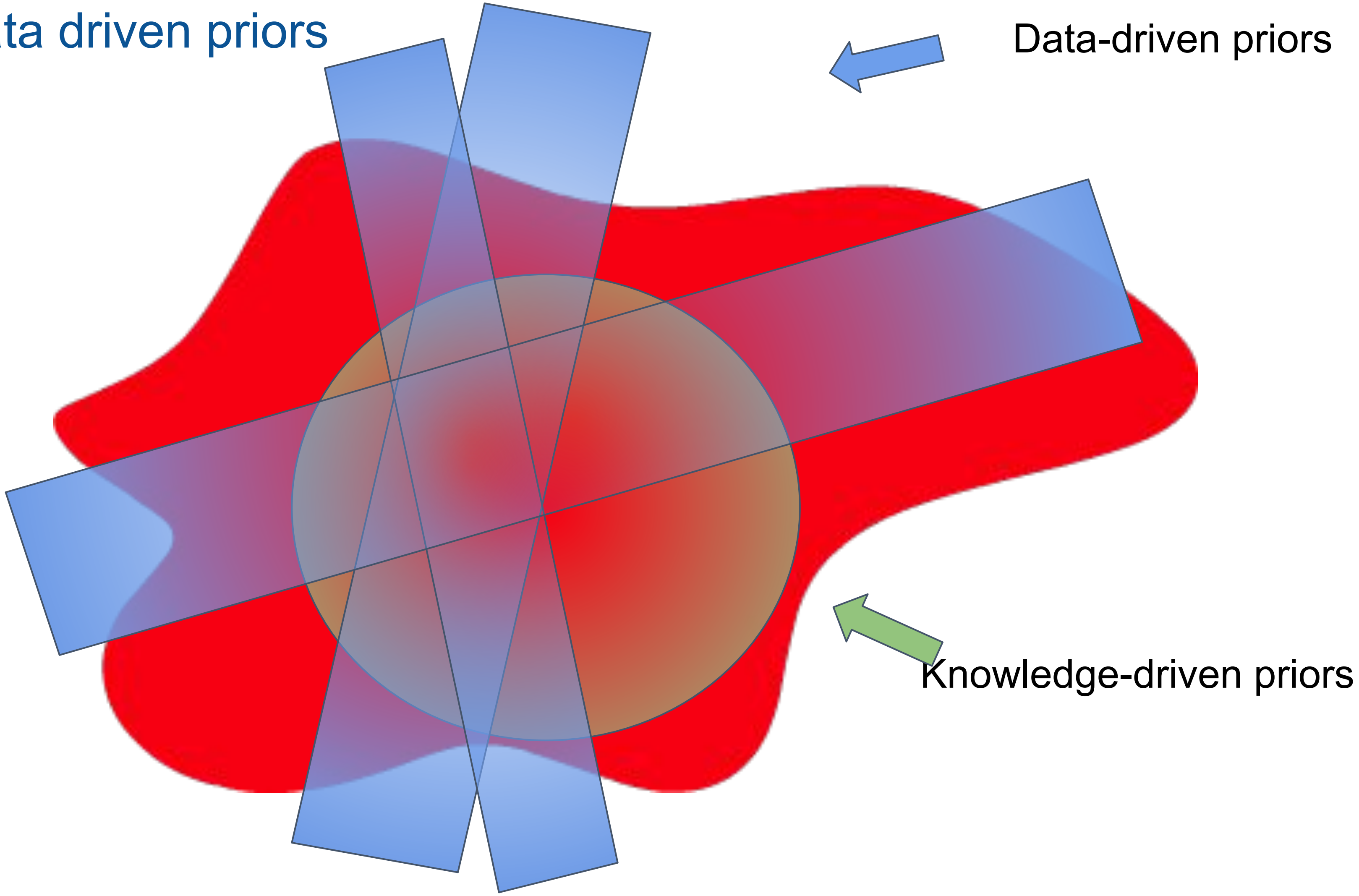
Constrained optimization (in CNNs)

Knowledge vs data driven priors



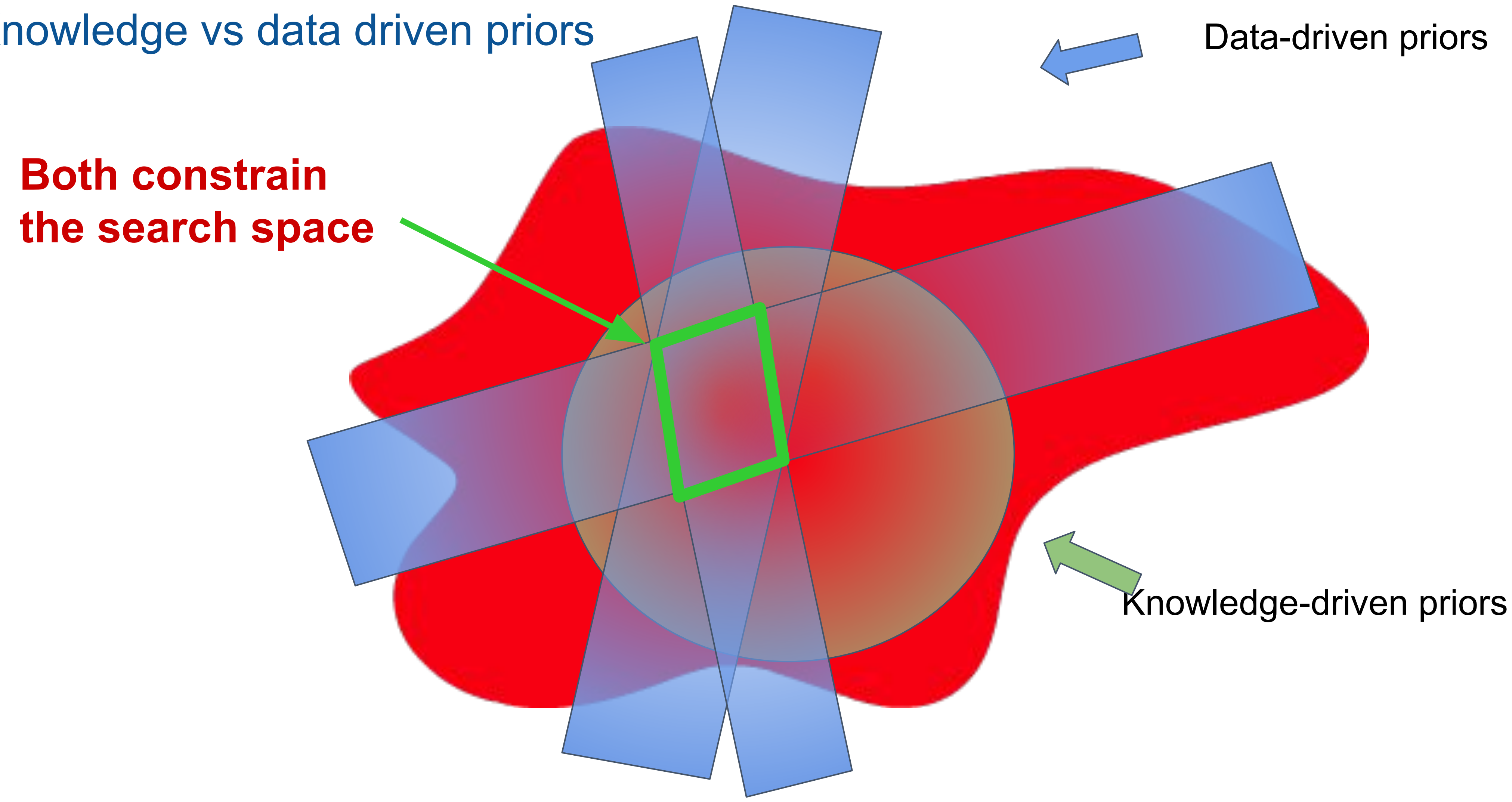
Constrained optimization (in CNNs)

Knowledge vs data driven priors



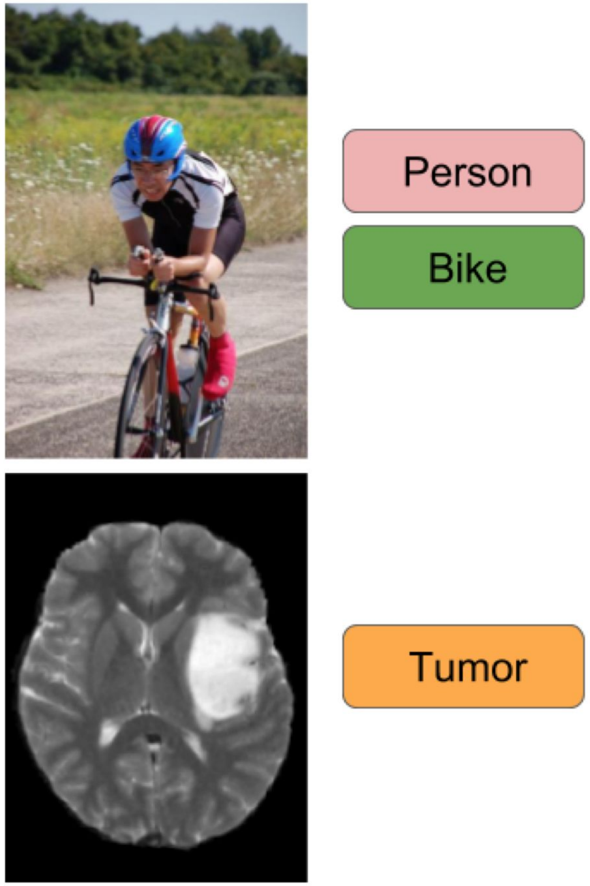
Constrained optimization (in CNNs)

Knowledge vs data driven priors



Data-driven priors (cues)

Image tags



Original Image

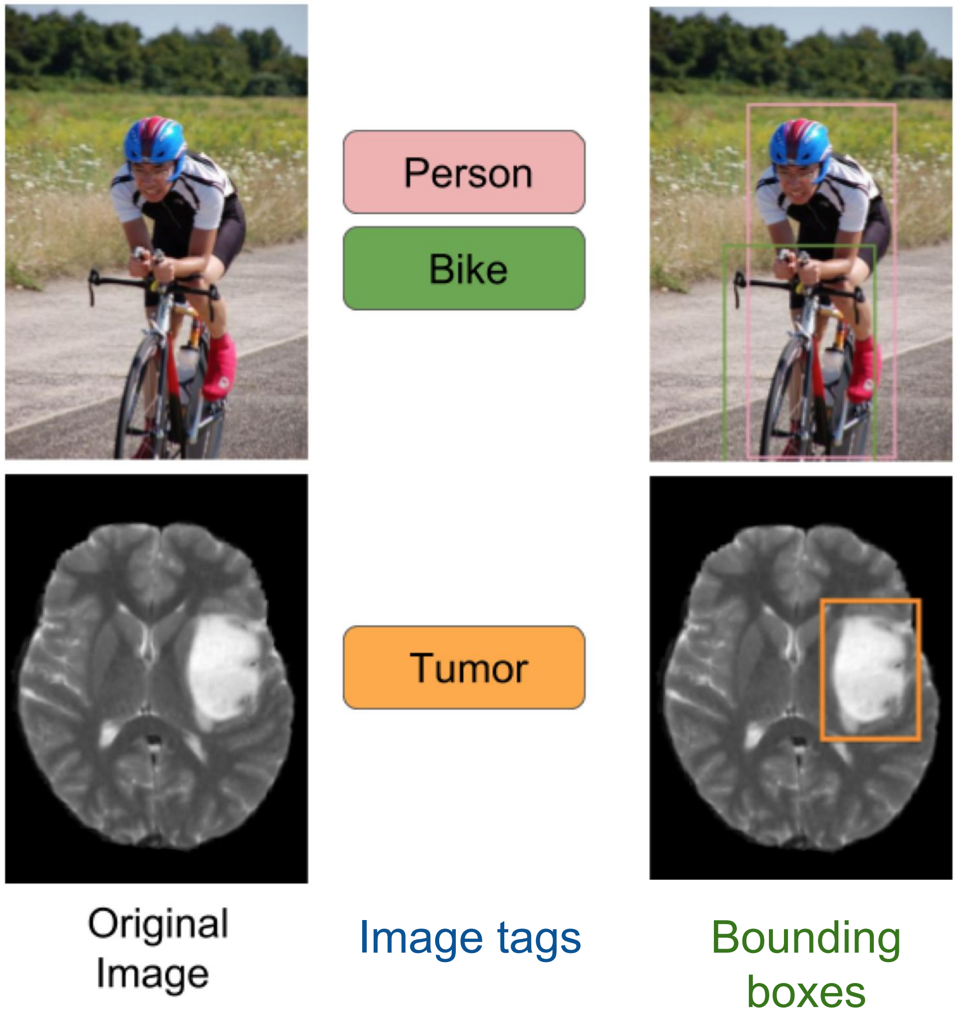
Image tags

- Pathak et al., Constrained convolutional neural networks for weakly supervised segmentation, ICCV 2015
- Kervadec et al., Constrained-CNN losses for weakly supervised segmentation, MedIA 2019.

Data-driven priors (cues)

Image tags

Bounding boxes



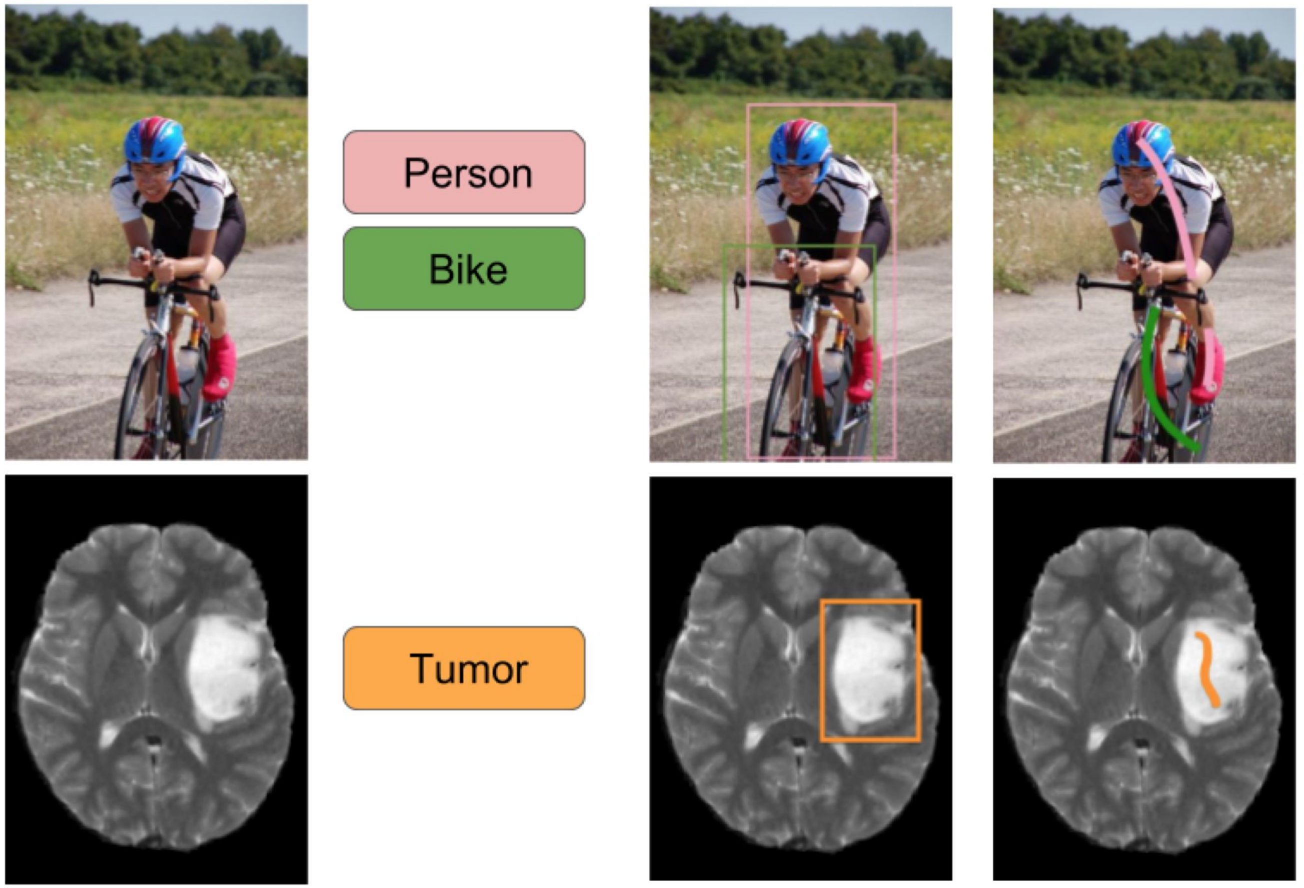
- Pathak et al., Constrained convolutional neural networks for weakly supervised segmentation, ICCV 2015
- Kervadec et al., Constrained-CNN losses for weakly supervised segmentation, MedIA 2019.

Data-driven priors (cues)

Image tags

Bounding boxes

Scribbles



Original Image | Image tags | Bounding boxes | Scribbles

- Pathak et al., Constrained convolutional neural networks for weakly supervised segmentation, ICCV 2015
- Kervadec et al., Constrained-CNN losses for weakly supervised segmentation, MedIA 2019.

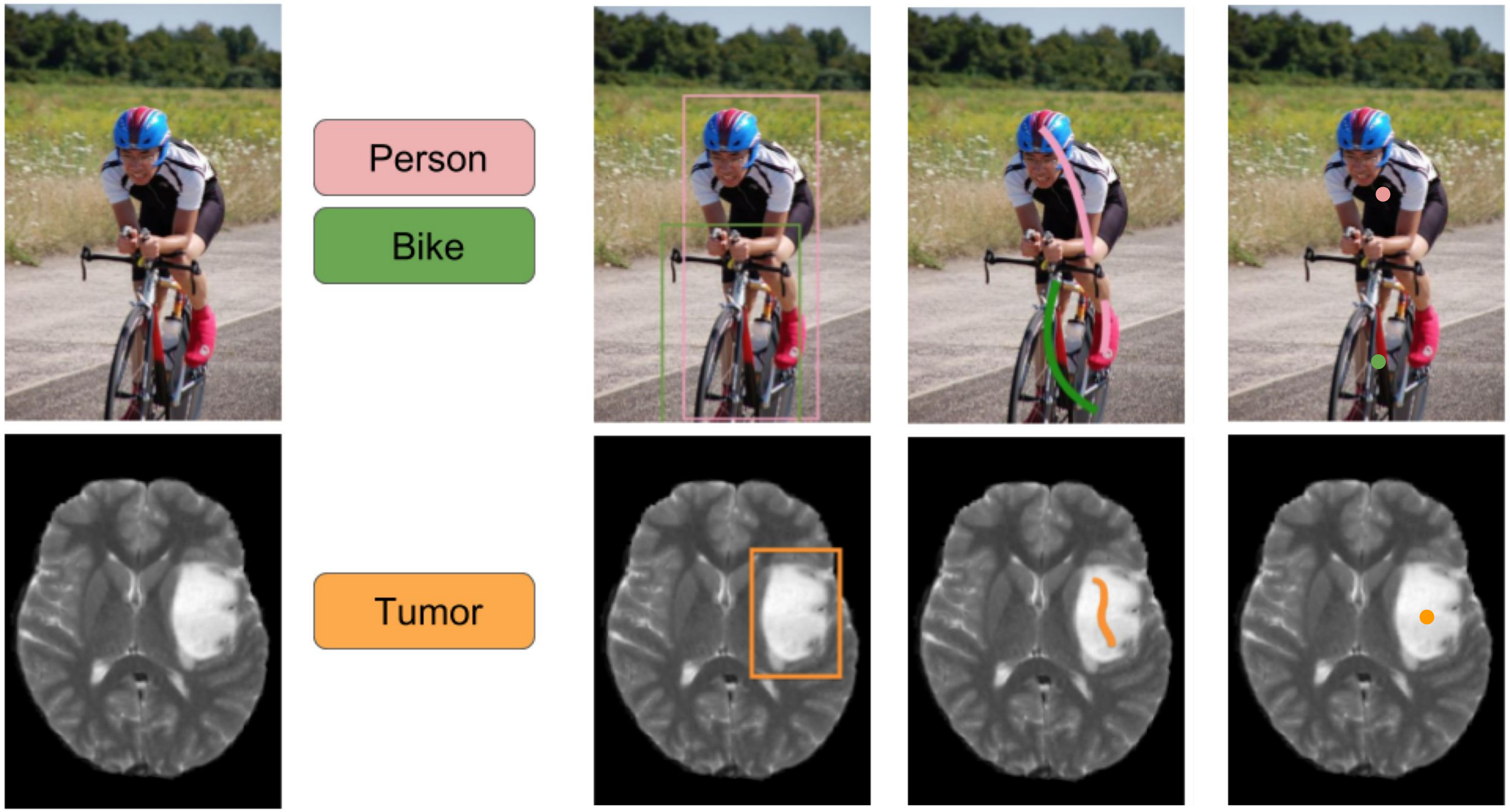
Data-driven priors (cues)

Image tags

Bounding boxes

Scribbles

Points



Original Image | Image tags | Bounding boxes | Scribbles | Points

- Pathak et al., Constrained convolutional neural networks for weakly supervised segmentation, ICCV 2015
- Kervadec et al., Constrained-CNN losses for weakly supervised segmentation, MedIA 2019.

Data-driven priors (cues)

Another data-driven priors

Image captions

A boy jumping on a skateboard

Extreme points

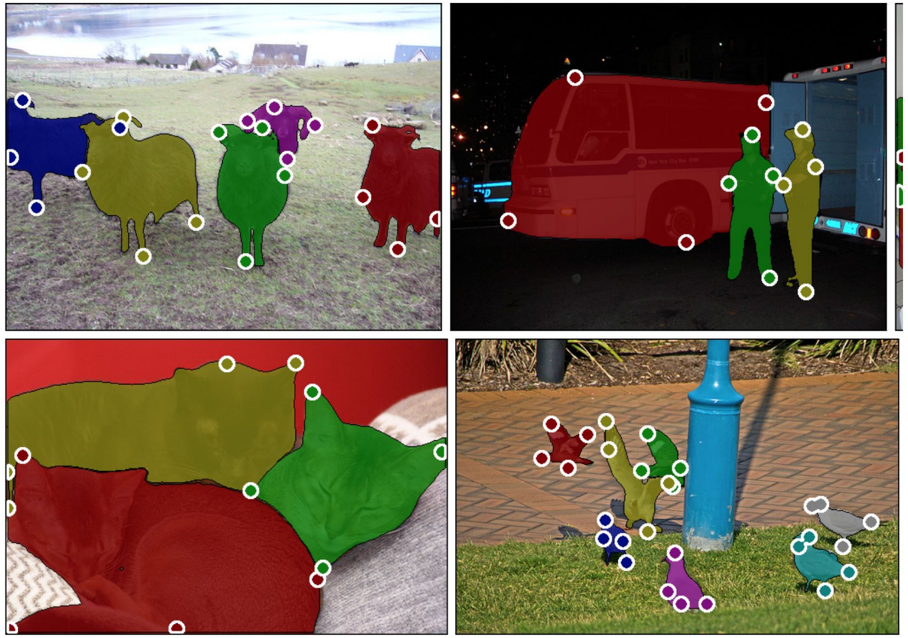
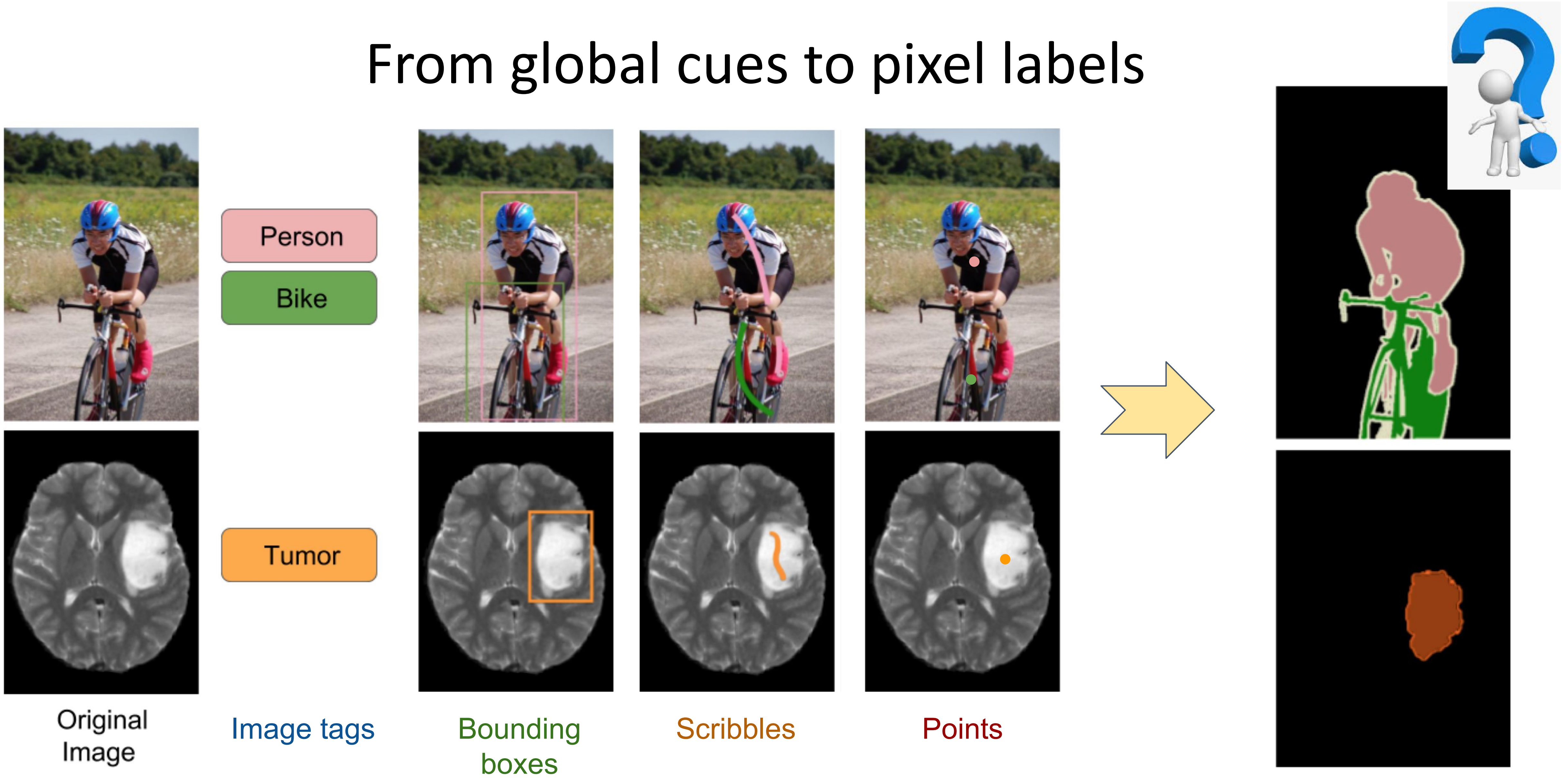


Image from Maninis et al, CVPR'18

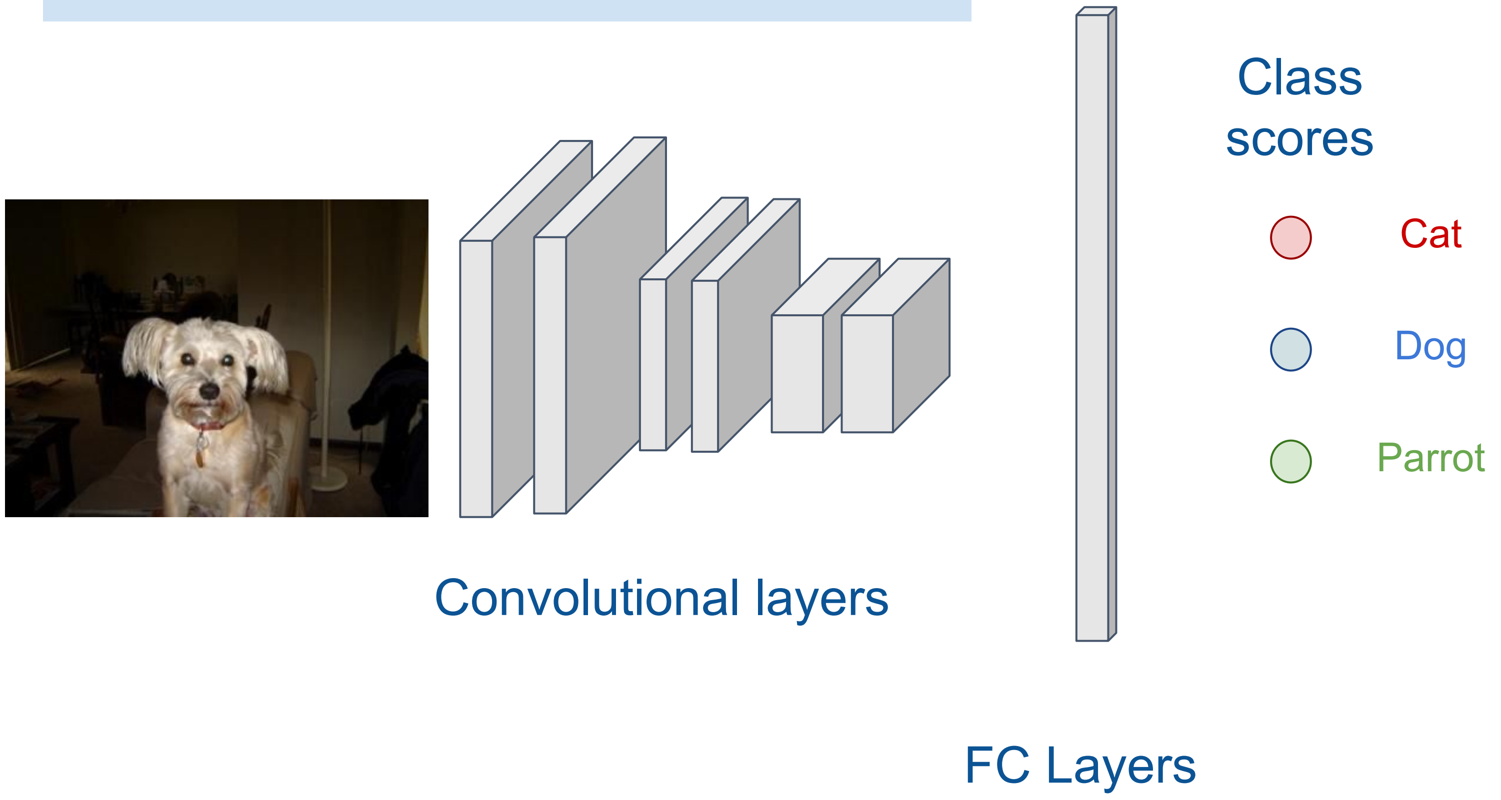
- Maninis et al. Deep extreme cut: From extreme points to object segmentation. CVPR 2018

From global cues to pixel labels



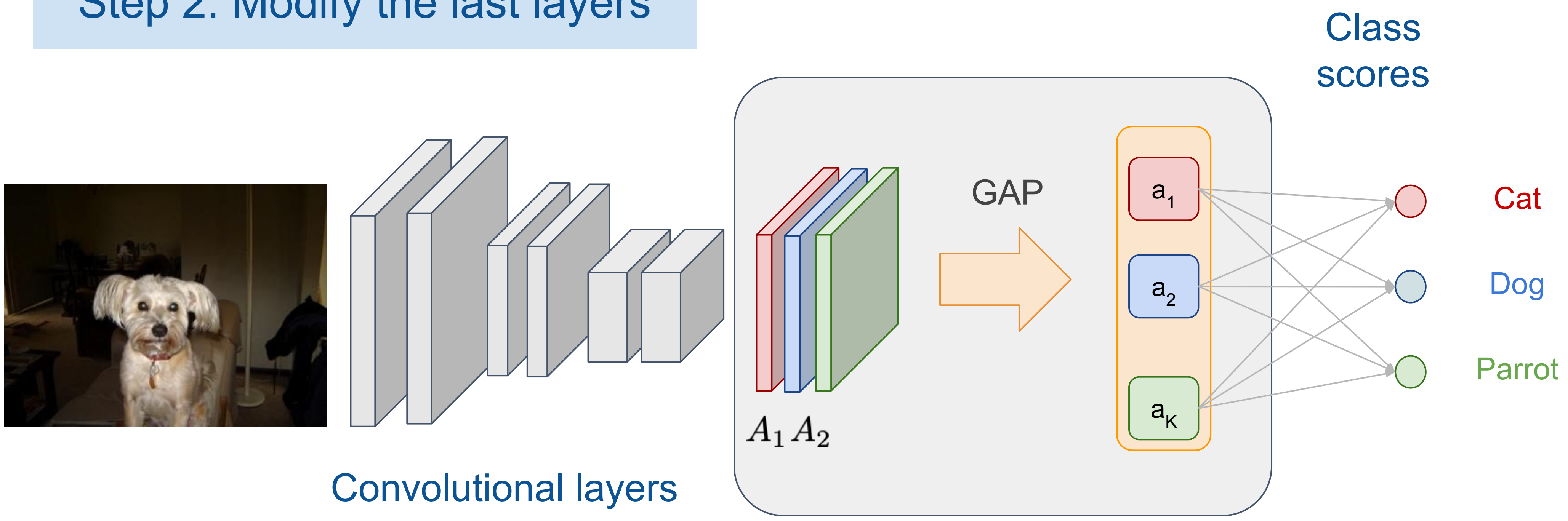
From global cues to pixel labels

Step 1: Get a classification CNN



From global cues to pixel labels

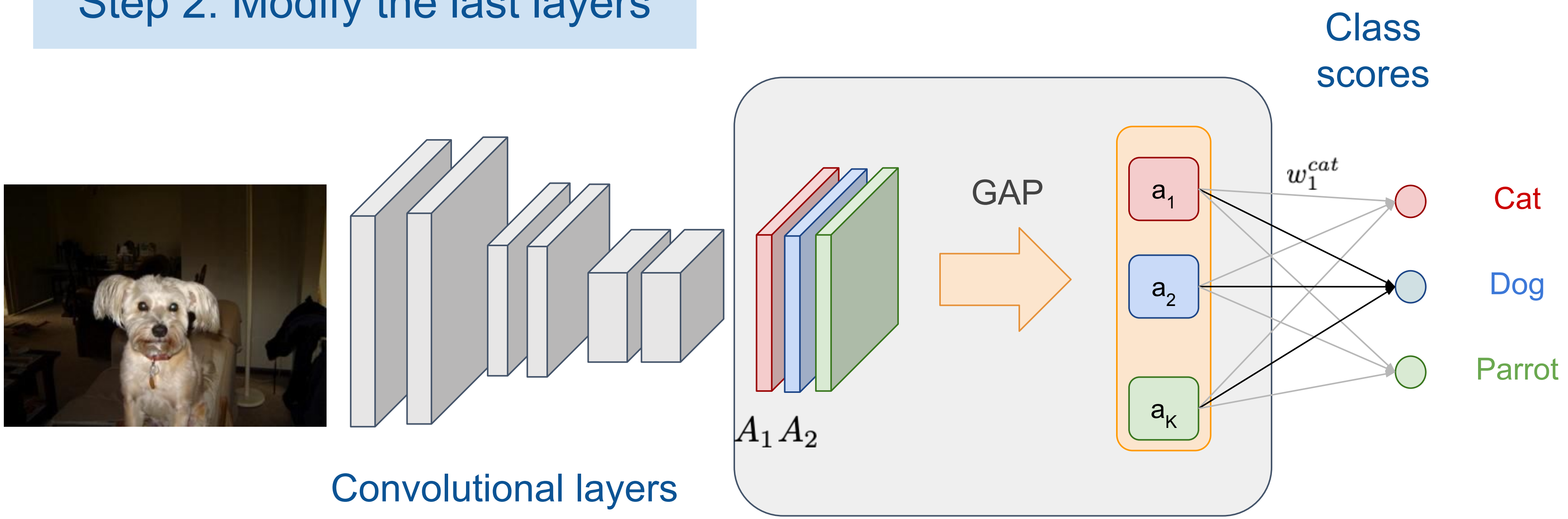
Step 2: Modify the last layers



$$GAP(A_k) = a_k = \frac{1}{|N|}\sum_{x,y} A_k(x,y)$$

From global cues to pixel labels

Step 2: Modify the last layers



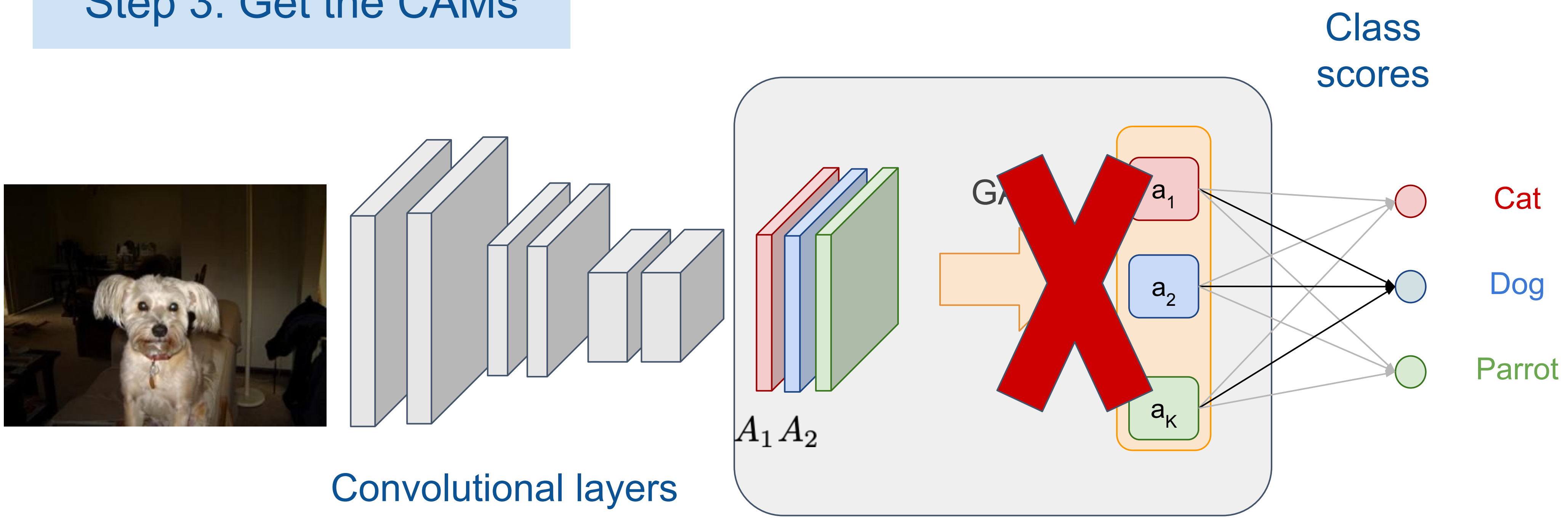
$$GAP(A_k) = a_k = \frac{1}{|N|}\sum_{x,y} A_k(x,y)$$

Class score (logits)

$$S_c = \sum_k w_k^c a_k = \frac{1}{N}\sum_k w_k^c \sum_{x,y} a_k(x,y)$$

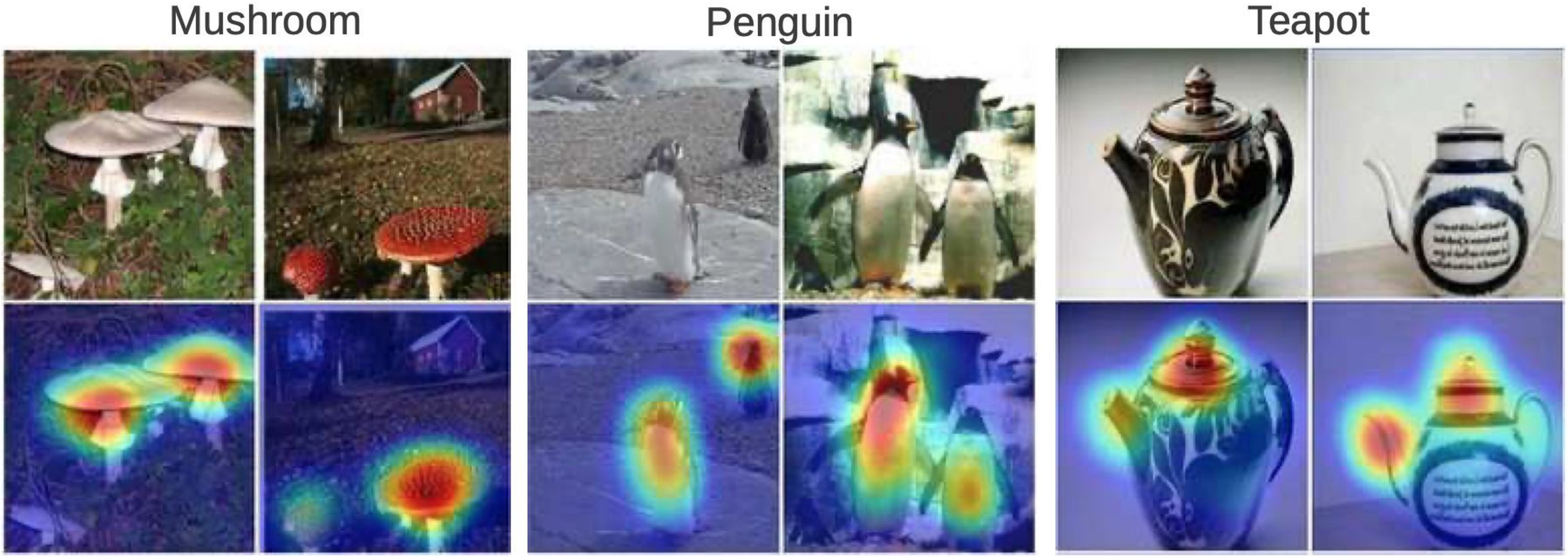
From global cues to pixel labels

Step 3: Get the CAMs



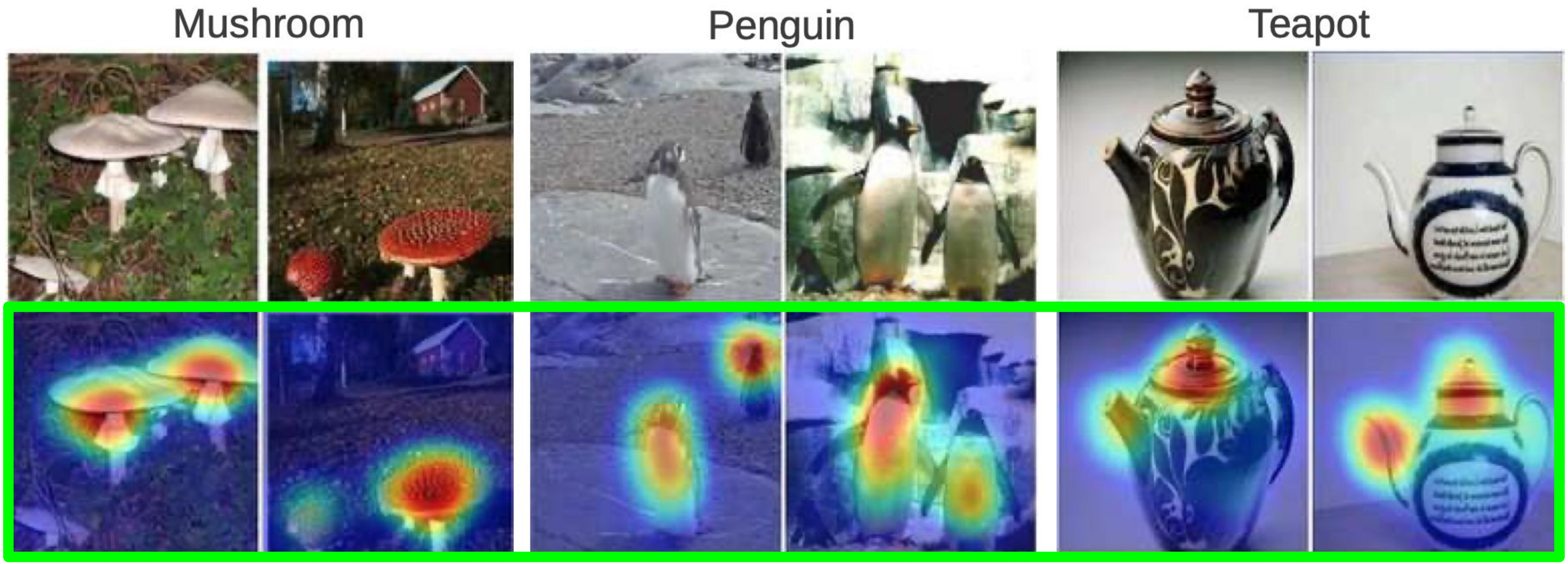
$$CAM_{Dog}(x,y) = \sum_k w_k^{Dog} A_k(x,y) =$$

From global cues to pixel labels



- Zhou et al., Learning deep features for discriminative localization. CVPR 2016

From global cues to pixel labels



These activations maps can be used as **pseudo-masks**

- Zhou et al., Learning deep features for discriminative localization. CVPR 2016

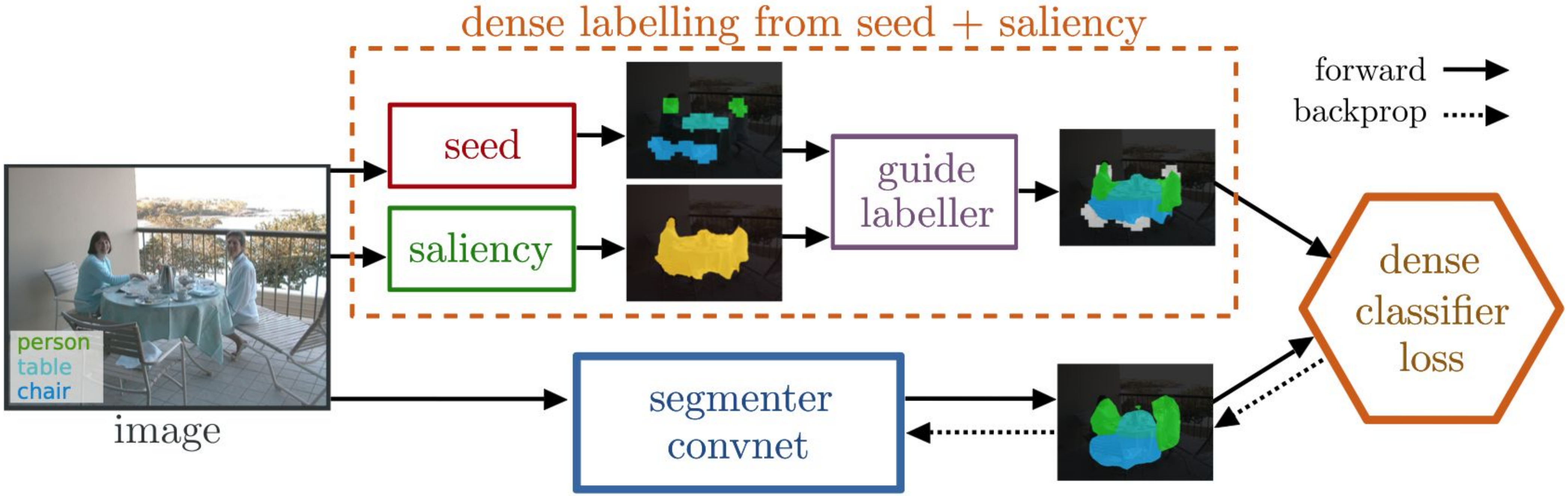
From global cues to pixel labels

Problem: they focus only on highly discriminative regions

From global cues to pixel labels

Problem: they focus only on highly discriminative regions

Incorporate saliency maps

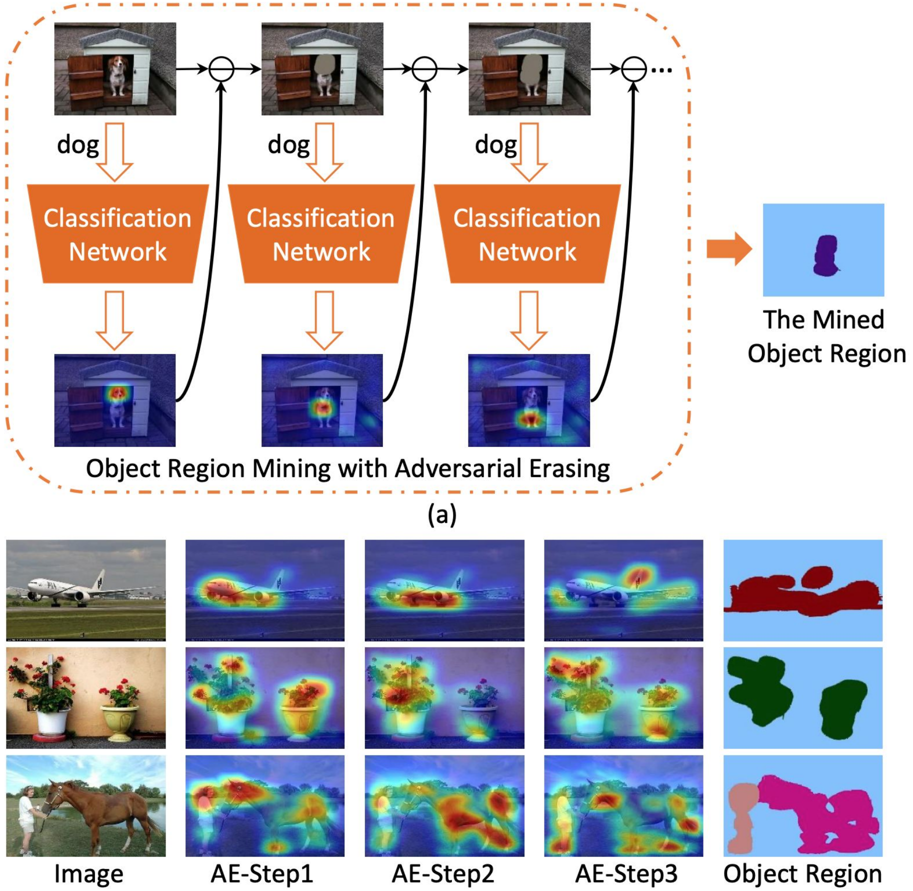


- Oh et al. Exploiting Saliency for Object Segmentation from Image Level Labels. CVPR 2017
- Fan et al. Learning Integral Objects With Intra-Class Discriminator for Weakly-Supervised Semantic Segmentation. CVPR 2020

From global cues to pixel labels

Problem: they focus only on highly discriminative regions

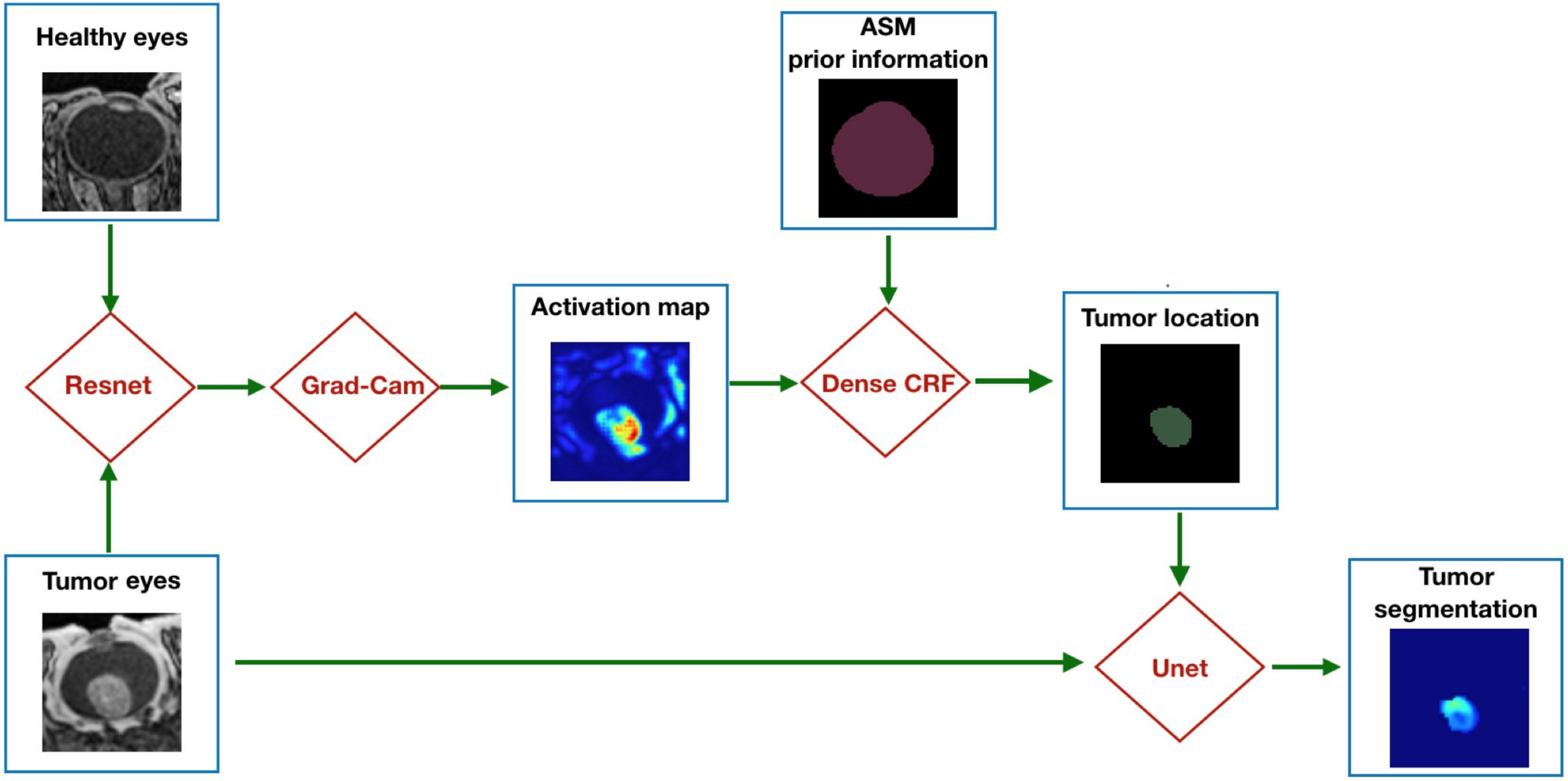
Region mining



- Wei et al. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. CVPR 2017
- Wang et al. Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features. CVPR 2018

From global cues to pixel labels

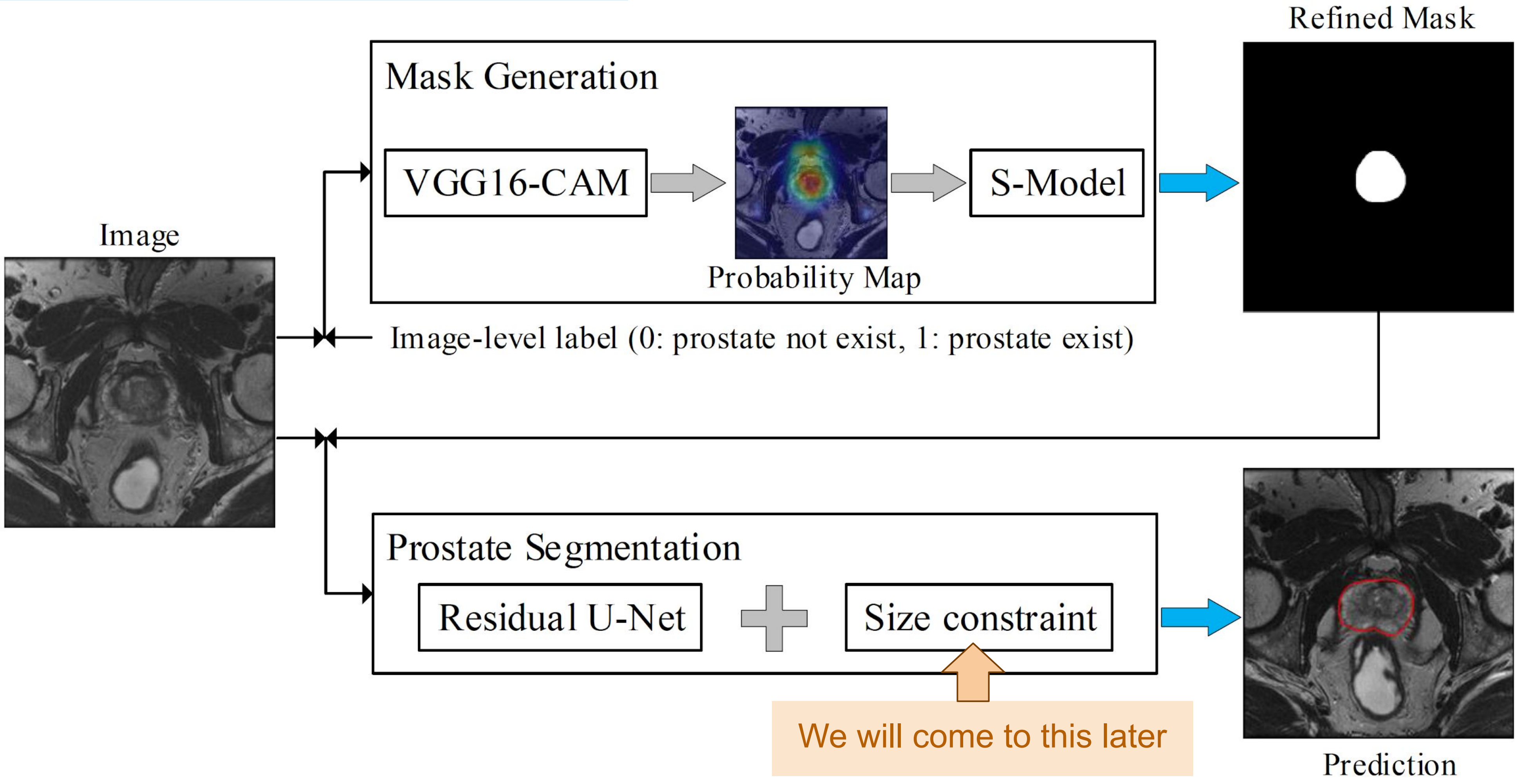
CAMs in the medical domain



- Nguyen et al. A novel segmentation framework for uveal melanoma based on magnetic resonance imaging and class activation maps. MIDL 2019.

From global cues to pixel labels

CAMs in the medical domain



We will come to this later

- Chen et al. Exploiting confident information for weakly supervised prostate segmentation based on image-level labels. SPIE Medical Imaging 2020

Constrained optimization (in CNNs)

Knowledge-driven priors

Common priors in natural images

Target Size

- Pathak et al., Constrained convolutional neural networks for weakly supervised segmentation, ICCV 2015
- Xu et al., Learning to Segment Under Various Forms of Weak Supervision, CVPR 2015
- Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV’17

Constrained optimization (in CNNs)

Knowledge-driven priors

Common priors in natural images

Target Location

- Remez et al. Learning to segment via cut-and-paste. ECCV 2018
- Georgakis et al Synthesizing training data for object detection in indoor scenes. RSS 2017

Constrained optimization (in CNNs)

Knowledge-driven priors

Common priors in natural images

Number of instances

- Deselaers et al. Localizing objects while learning their appearance. ECCV 2010

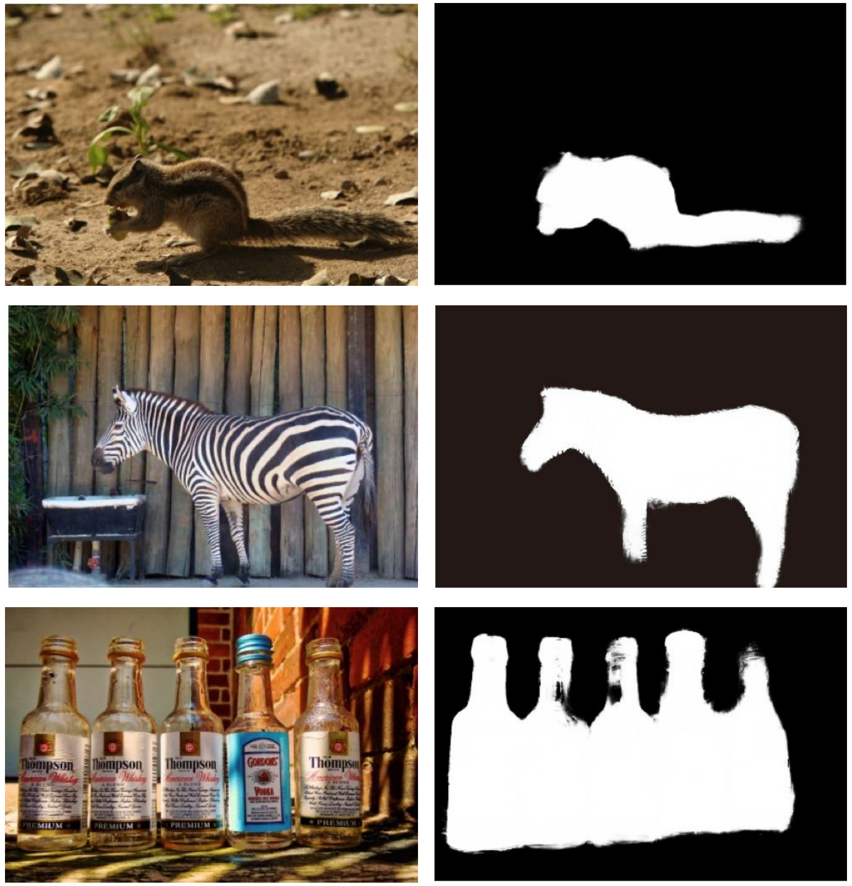
Constrained optimization (in CNNs)

Knowledge-driven priors

Common priors in natural images

Contrast
Foreground/Background

Saliency



Images from Hou et al, CVPR'17

- Hou et al. Deeply supervised salient object detection with short connections. CVPR 2017
- Li et al. Instance-level salient object segmentation. CVPR 2017

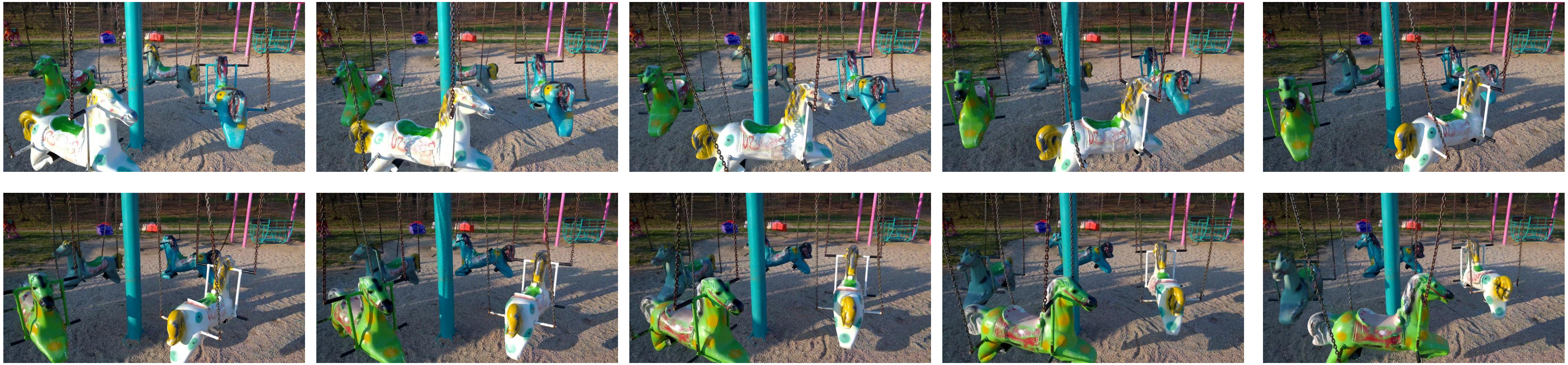
Constrained optimization (in CNNs)

Knowledge-driven priors

Common priors in natural images

Motion

Images from the DAVIS Challenge Dataset

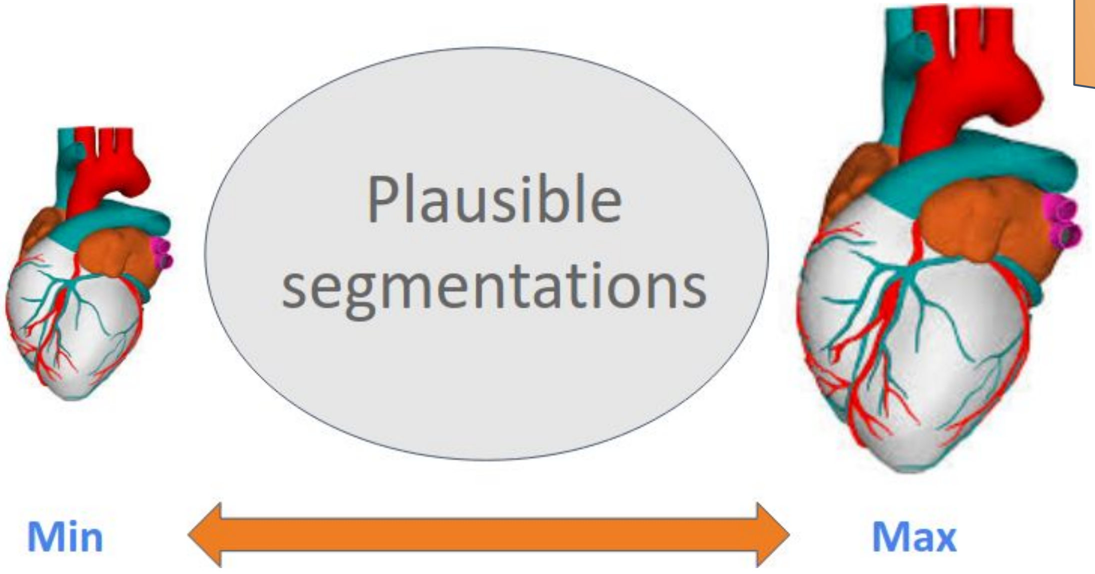


- Tokmakov et al. Weakly-supervised semantic segmentation using motion cues. ECCV 2016
- Pathak et al. Learning features by watching objects move. CVPR 2017

Constrained optimization (in CNNs)

Knowledge-driven priors

What about priors in the medical domain?

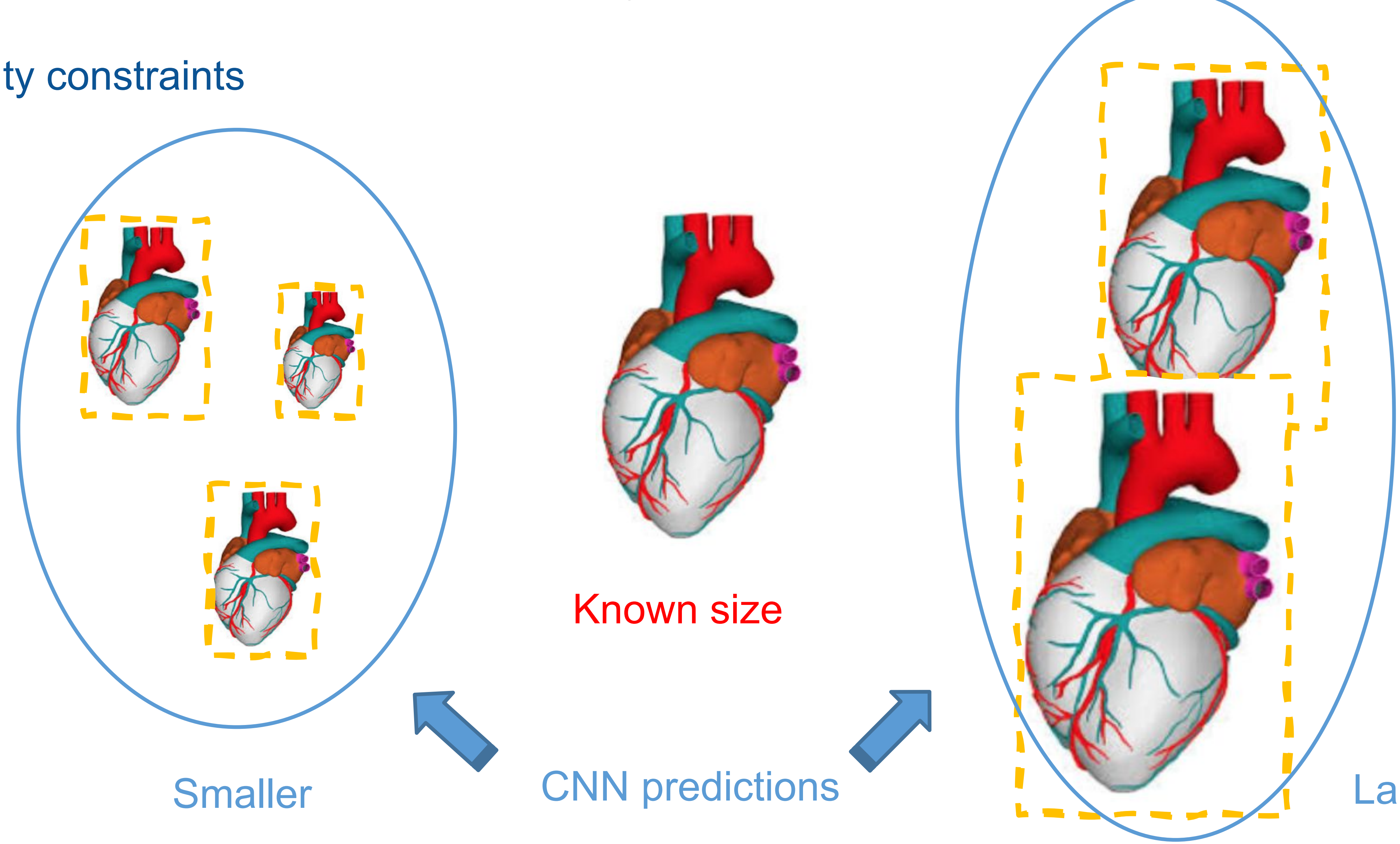


Anatomical priors

Partial labeled data
(exploit target relationships)

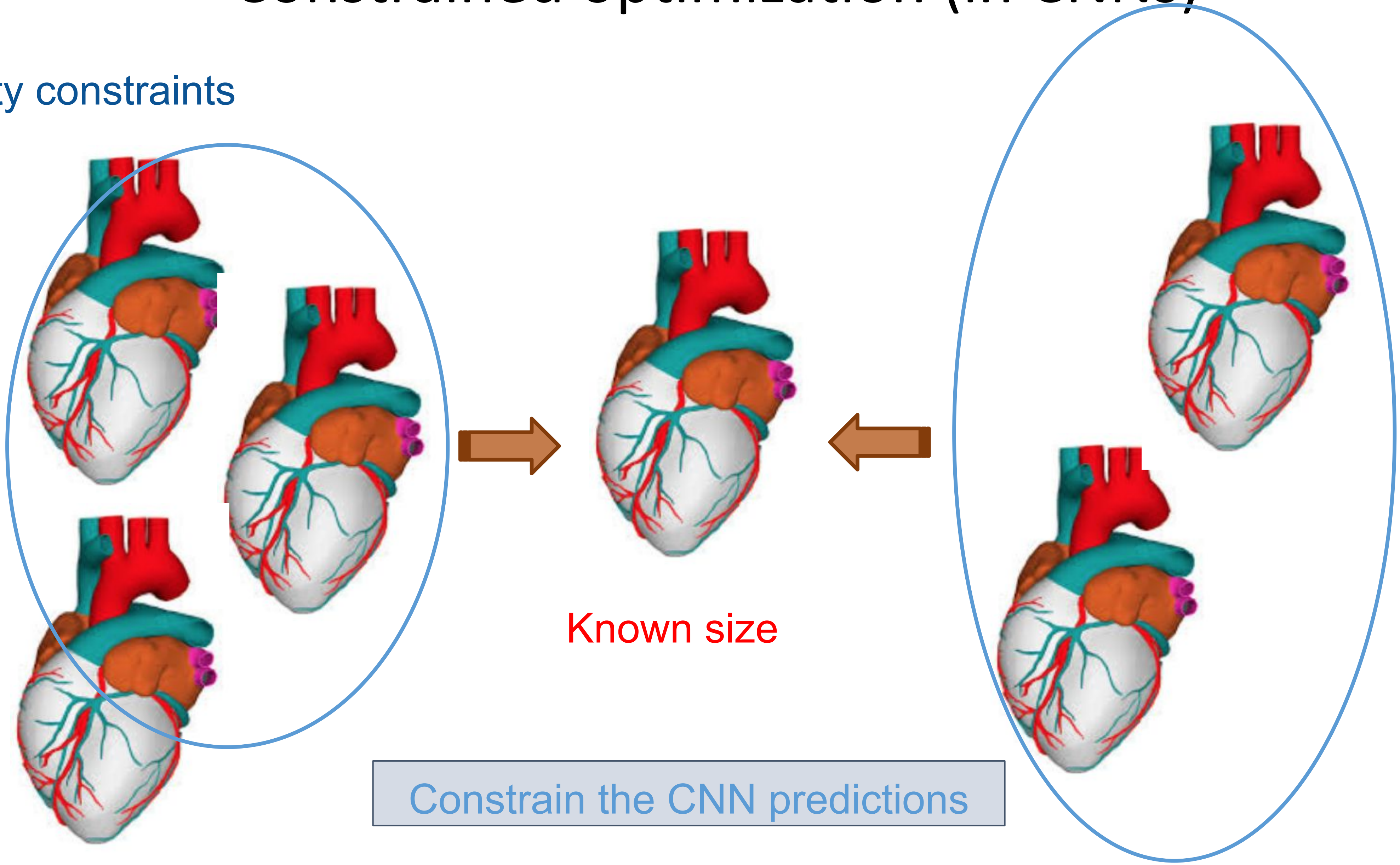
Constrained optimization (in CNNs)

Equality constraints



Constrained optimization (in CNNs)

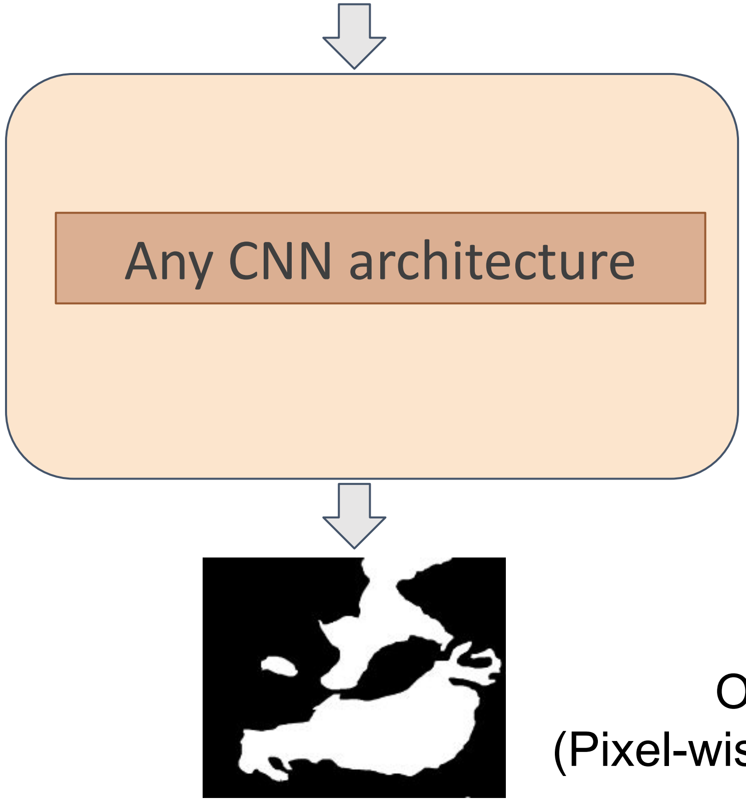
Equality constraints



Constrained optimization (in CNNs)

Equality constraints (e.g, L2 penalty)

Input
(Histology image)



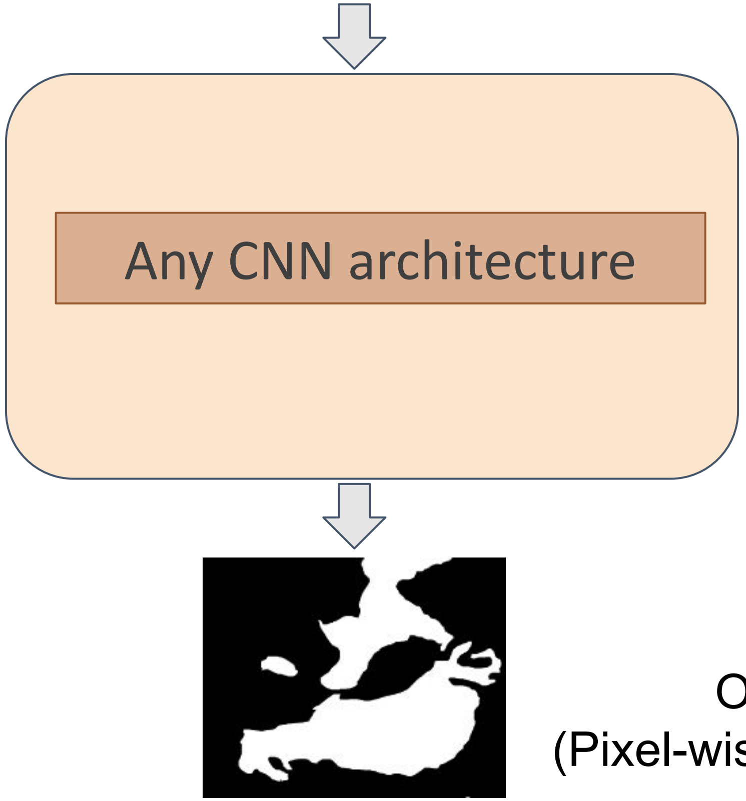
Output
(Pixel-wise prediction)

Jia et al. Constrained deep weak supervision for histopathology image segmentation. IEEE TMI 2017

Constrained optimization (in CNNs)

Equality constraints (e.g, L2 penalty)

Input
(Histology image)



Output
(Pixel-wise prediction)

Additional term

Reminder: ½ KL is an upper bound on L_2 for simplex vectors (Pinsker's inequality)

$$\boxed{l_{ac} = \boldsymbol{I}(Y_i = 1) \sum_i (v_i - a_i)^2}$$

Predicted relative size (%)

$$v_i = \frac{1}{N} \sum_{p \in \Omega} s_\theta^{p,1}$$

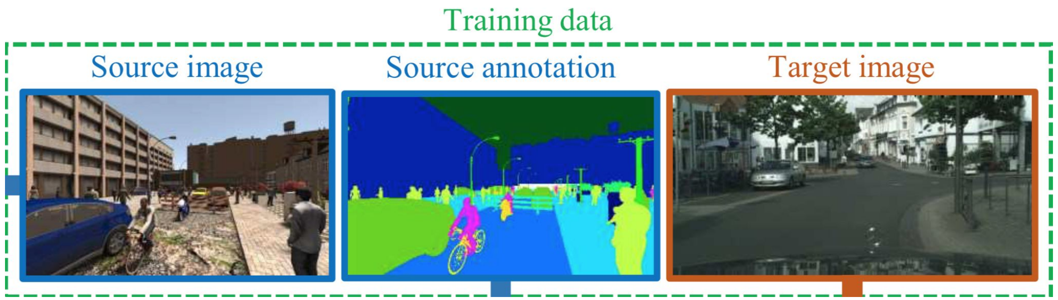
Predicted size given by experts

Jia et al. Constrained deep weak supervision for histopathology image segmentation. IEEE TMI 2017

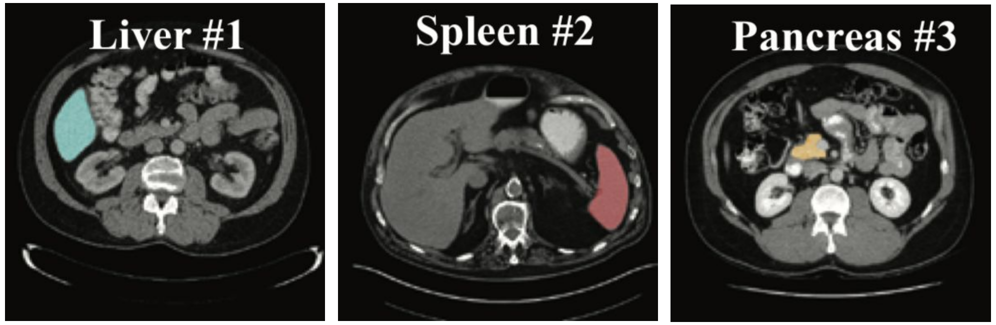
Constrained optimization (in CNNs)

Equality constraints (e.g, KL)

Unsupervised domain adaptation

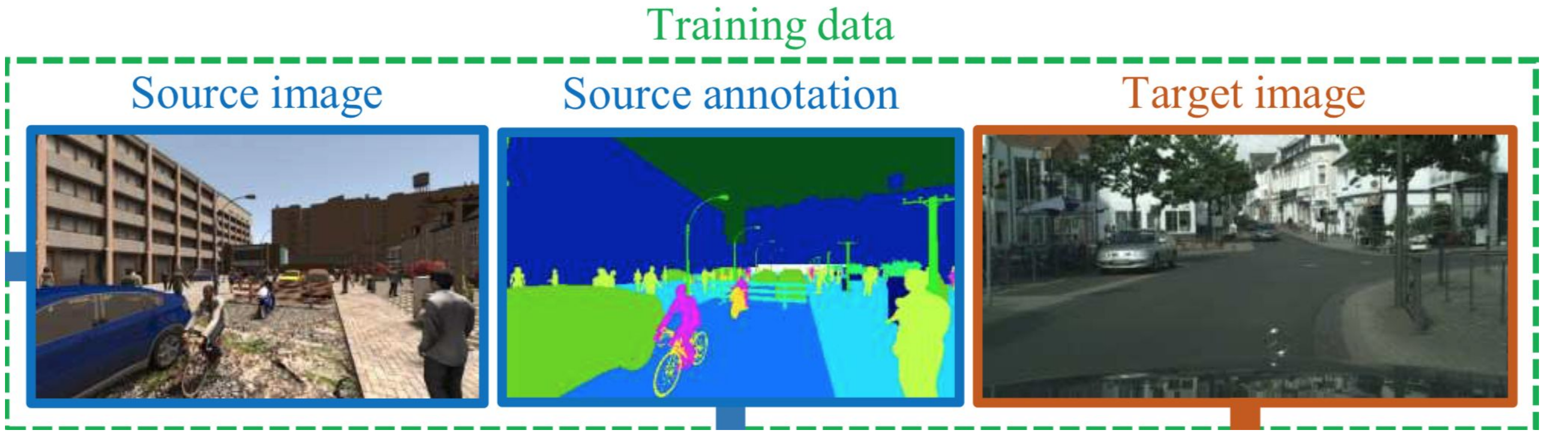


Partially labeled data



Constrained optimization (in CNNs)

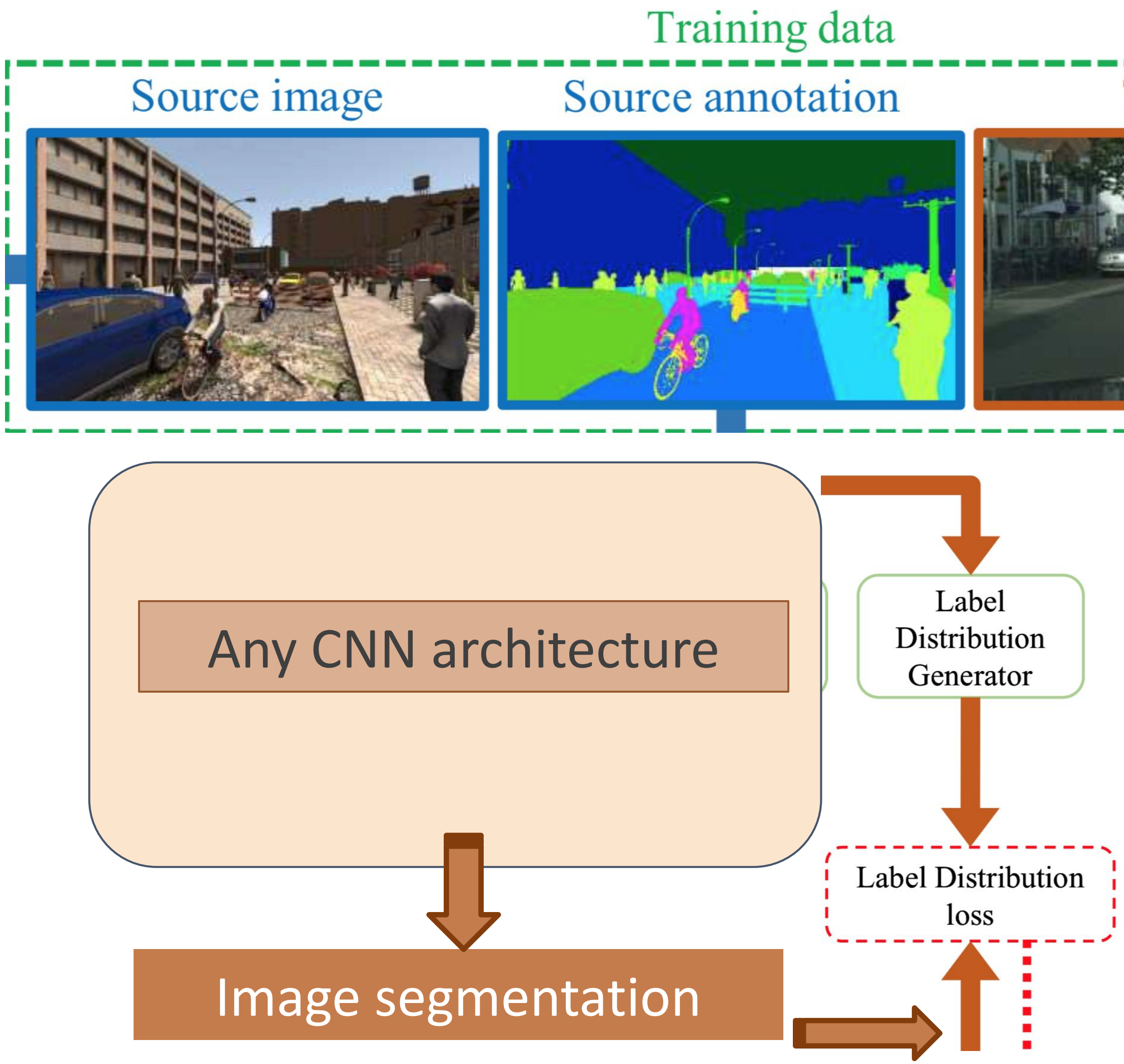
Equality constraints (e.g, KL): Curriculum DA



Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA

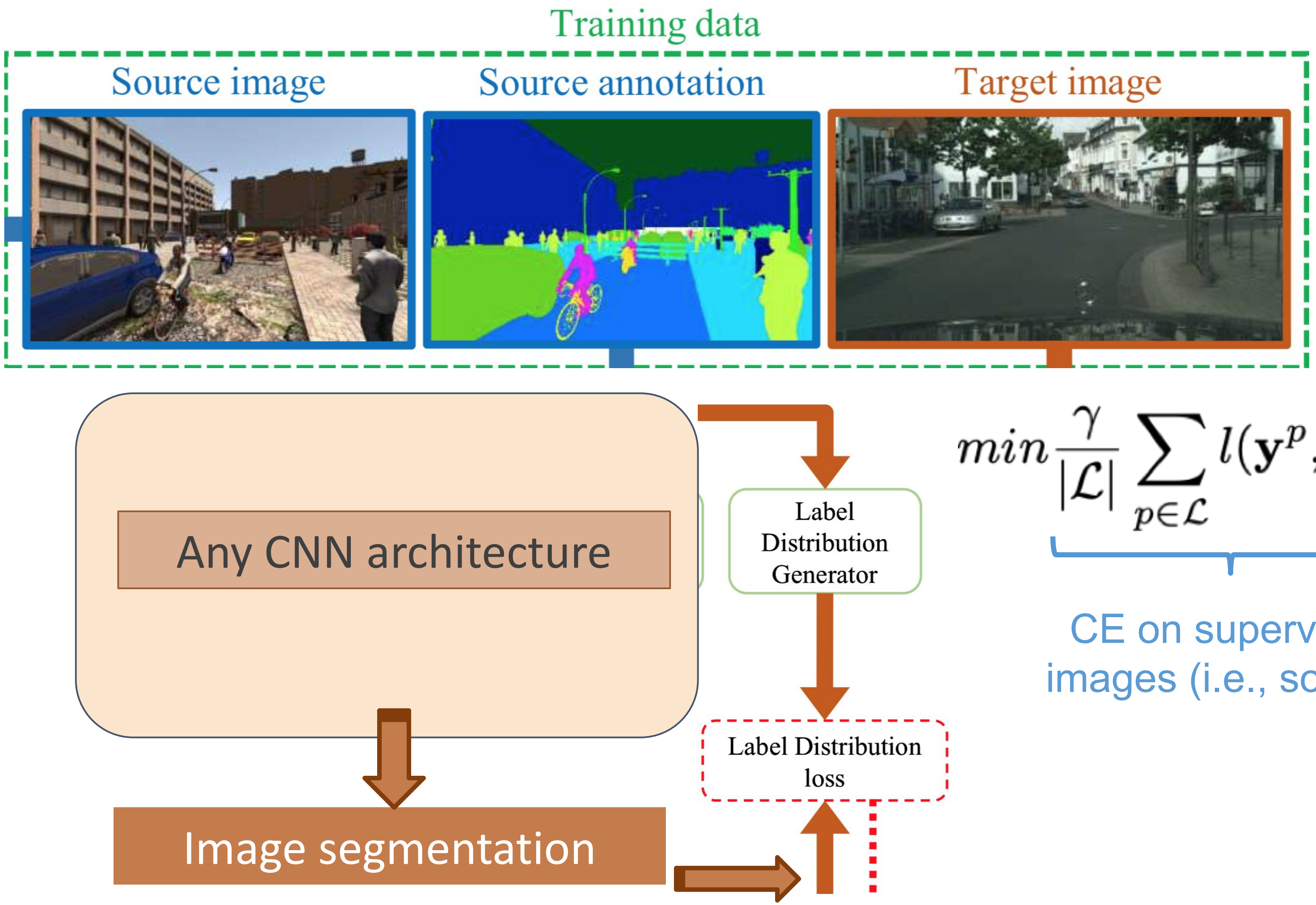


$$min \frac{\gamma}{|\mathcal{L}|}\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s^p_\theta) + \frac{1-\gamma}{|\mathcal{U}|}\sum_{q\in\mathcal{U}}\sum_{k} \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA



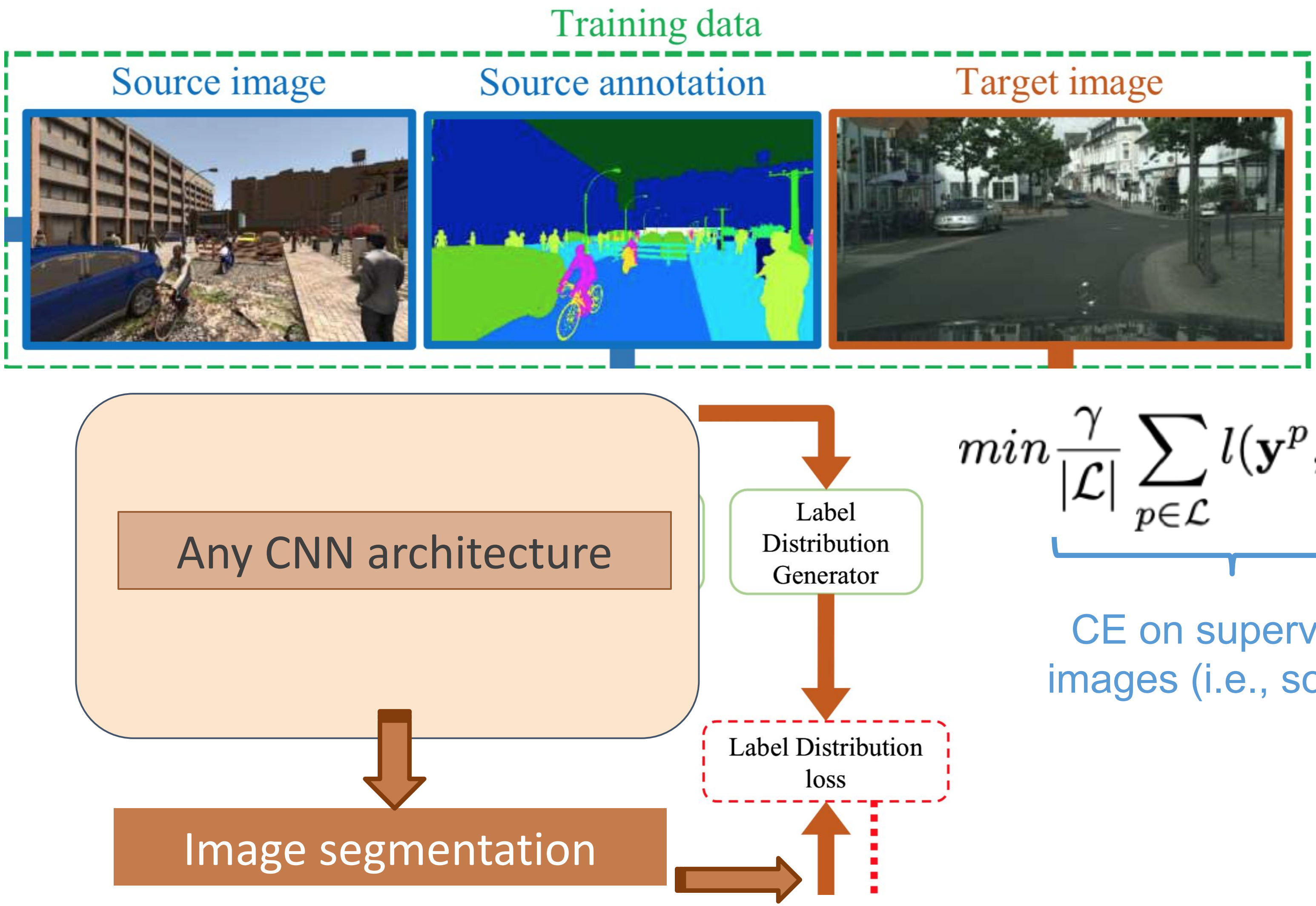
$$\min \underbrace{\frac{\gamma}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p)}_{\text{CE on supervised images (i.e., source)}} + \frac{1-\gamma}{|\mathcal{U}|} \sum_{q \in \mathcal{U}} \sum_k \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

CE on supervised images (i.e., source)

Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA



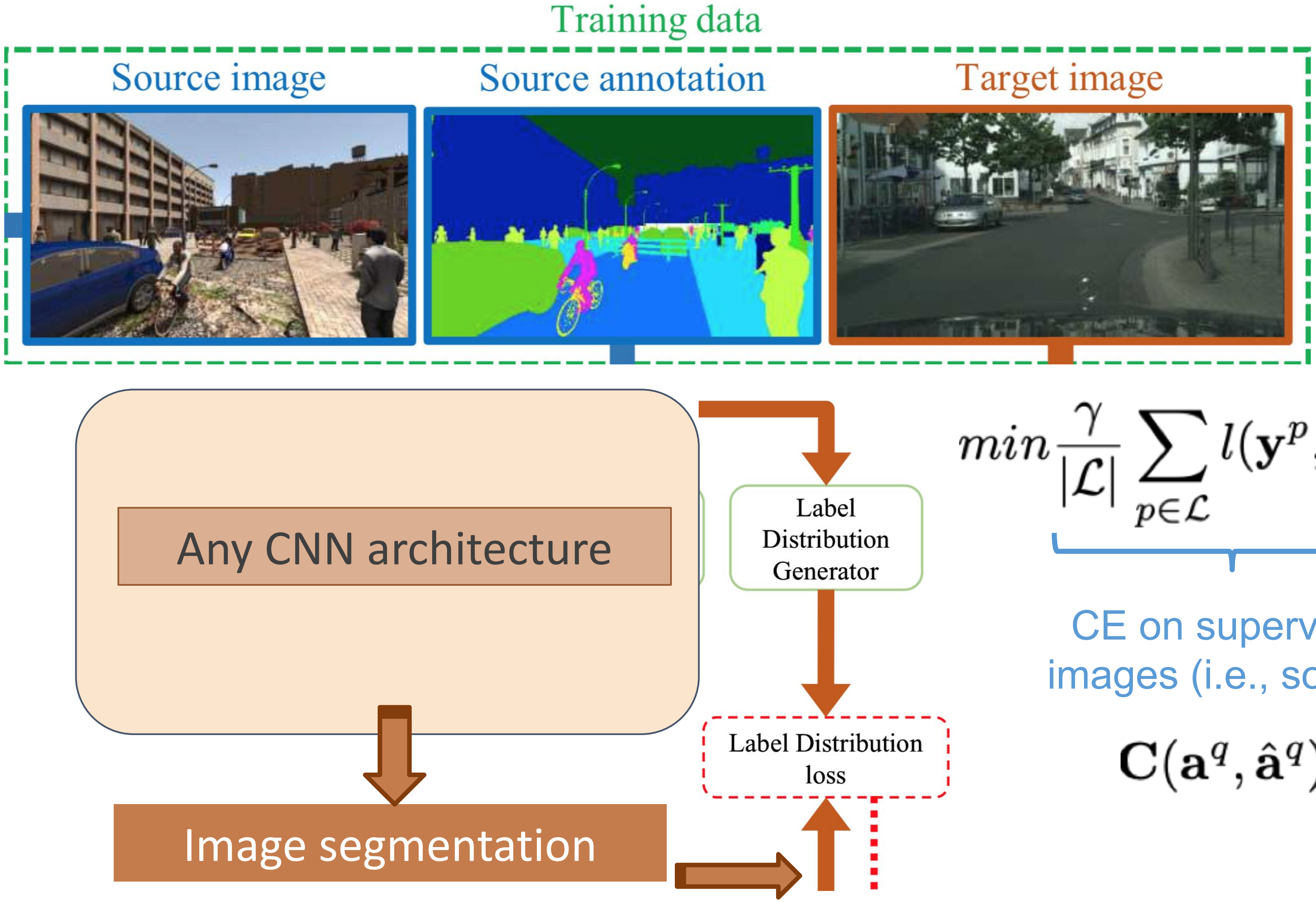
Additional term

$$\min \underbrace{\frac{\gamma}{|\mathcal{L}|}\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s_\theta^p)}_{\text{CE on supervised images (i.e., source)}} + \frac{1-\gamma}{|\mathcal{U}|}\sum_{q\in\mathcal{U}}\sum_{k} \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA



Additional term

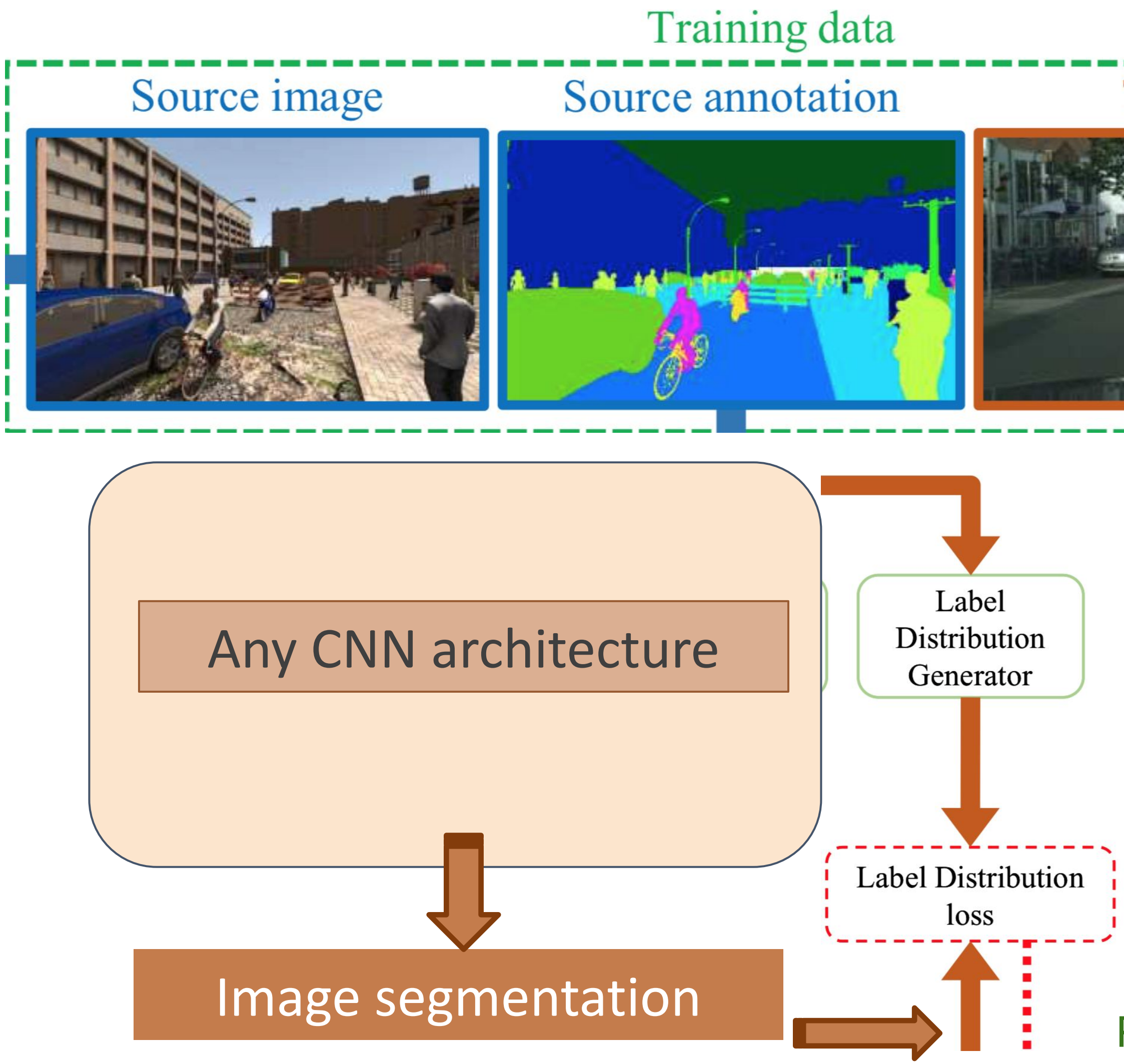
$$min \underbrace{\frac{\gamma}{|\mathcal{L}|}\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s^p_\theta)}_{\text{CE on supervised images (i.e., source)}} + \frac{1-\gamma}{|\mathcal{U}|}\sum_{q\in\mathcal{U}}\sum_{k} \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

$$\mathbf{C}(\mathbf{a}^q, \hat{\mathbf{a}}^q) = H(\mathbf{a}^q) + KL(\mathbf{a}^q, \hat{\mathbf{a}}^q)$$

Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA



Additional term

$$\min \frac{\gamma}{|\mathcal{L}|}\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s_\theta^p) + \frac{1-\gamma}{|\mathcal{U}|}\sum_{q\in\mathcal{U}}\sum_{k} \mathbf{C}(a^{q,k}, \hat{a}^{q,k})$$

CE on supervised images (i.e., source)

$$\mathbf{C}(\mathbf{a}^q, \hat{\mathbf{a}}^q) = H(\mathbf{a}^q) + KL(\mathbf{a}^q, \hat{\mathbf{a}}^q)$$

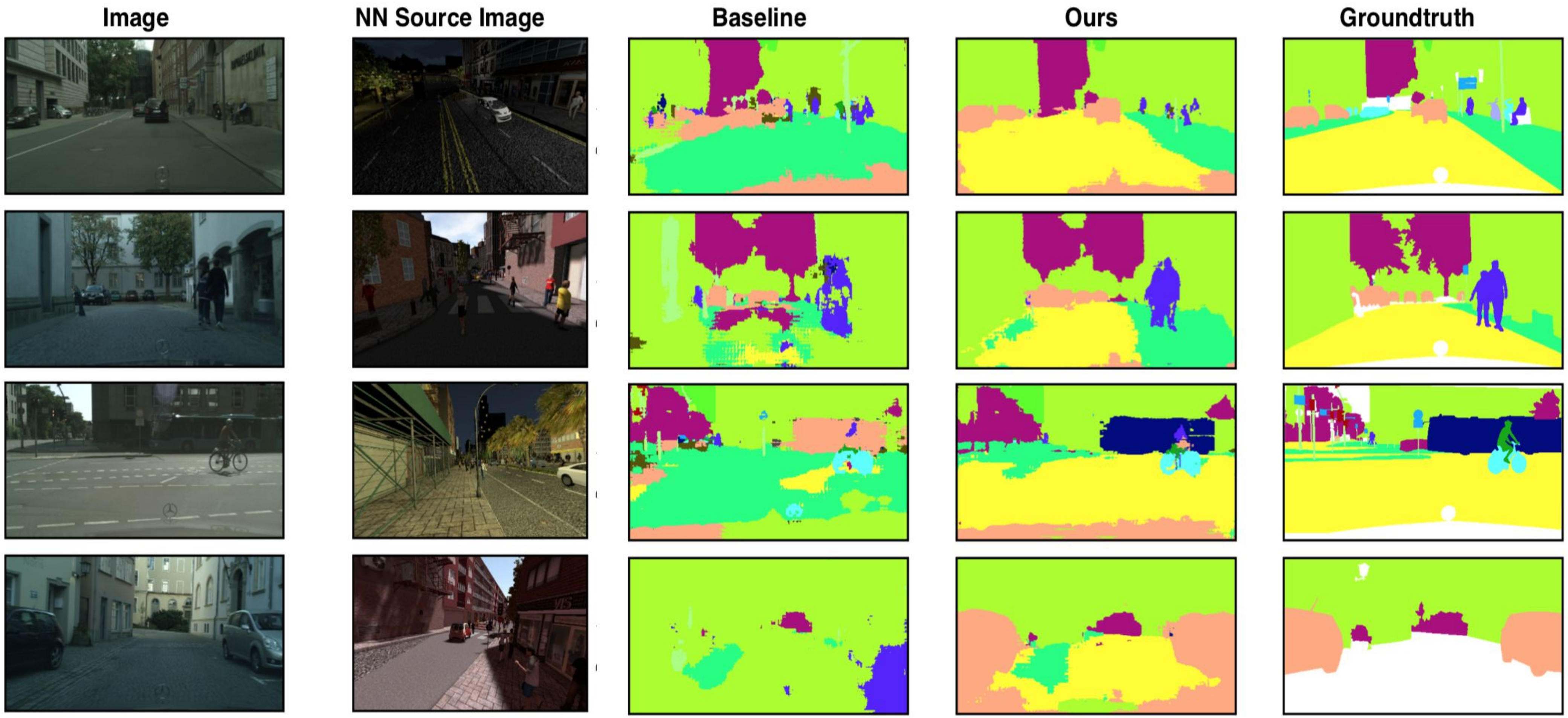
Predicted size

From predicted image

Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV'17

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Curriculum DA



Images from [Zhang et al., Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. ICCV’17]

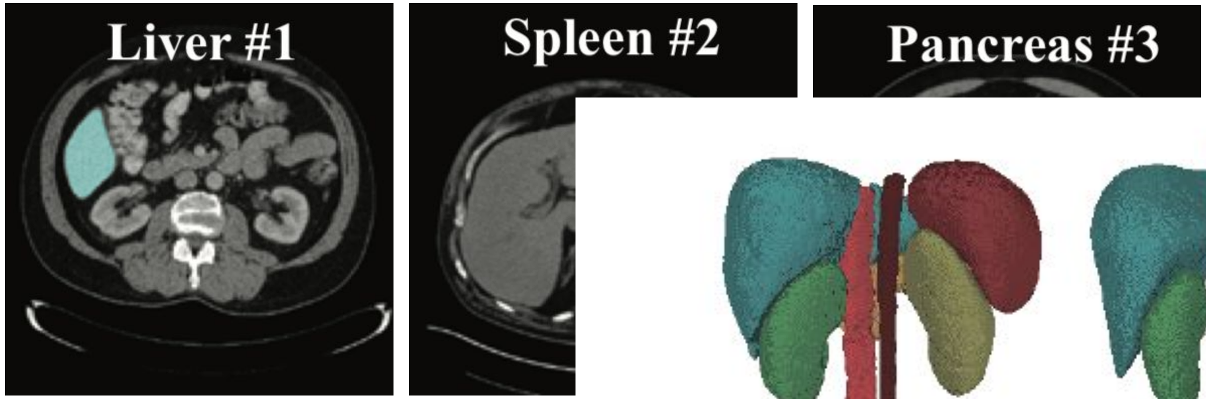
Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations



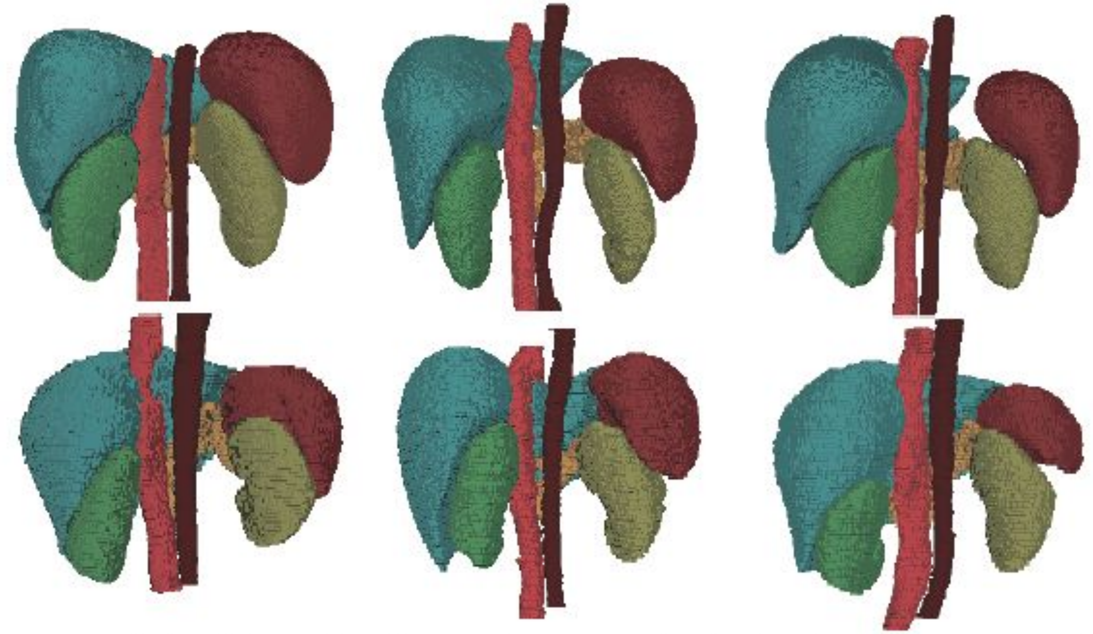


Figure 1. 3D Visualization of several abdominal organs (liver, spleen, left kidney, right kidney, aorta, inferior vena cava) to show the similarity of patient-wise abdominal organ size distributions.

Prior on the proportion

Images from [Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Main objective:

$$min \frac{1}{|\mathcal{L}|} \sum_{p \in \mathcal{L}} l(\mathbf{y}^p, \mathbf{s}_\theta^p) + \lambda_1 \frac{1}{|\mathcal{P}|} \sum_{q \in \mathcal{P}} l(\mathbf{y}^q, \mathbf{s}_\theta^q) + \lambda_2 \mathcal{J}(\theta)$$

Prior-aware loss

Fully labeled images

Partially labeled images

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Prior-aware loss

Averaged predicted distribution

$$\hat{\mathbf{p}} = \frac{1}{N} \sum_{p \in \mathcal{P}} \mathbf{s}_\theta^p$$

$$[s_\theta^{p,0}, s_\theta^{p,1}, \ldots, s_\theta^{p,|K|}]$$

On partially labeled images

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Prior-aware loss

Embed prior knowledge

$$KL(\mathbf{q}|\hat{\mathbf{p}})$$

Real label distribution   Average predicted distribution

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV’19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

KL can be expanded

$$\sum_c KL(q^c|\hat{p}^c) = -\sum_c (q^c \log \hat{p}^c + ((1-q^c)\log(1-\hat{p}^c)) + const$$

Prior-aware loss

$$-\sum_{c=0}^{|K|}\{q^c log \frac{1}{N}\sum_{p\in\mathcal{P}} \mathbf{s}_\theta^{p,c} + (1-q^c) log(1-\frac{1}{N}\sum_{p\in\mathcal{P}} \mathbf{s}_\theta^{p,c})\} + const$$

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Prior-aware loss

KL can be expanded

$$-\sum_{c=0}^{|K|}\{q^c log \frac{1}{N}\sum_{p\in\mathcal{P}} \mathbf{s}_\theta^{p,c} + (1-q^c)log(1-\frac{1}{N}\sum_{p\in\mathcal{P}} \mathbf{s}_\theta^{p,c})\} + const$$

This is problematic (average distribution of p̂ organ sizes inside log!!)

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

Stochastic primal-dual gradient

(split terms updated independently)

Prior-aware loss

KL can be expanded

$$-\sum_{c=0}^{|K|}\{q^c log \frac{1}{N}\sum_{p\in\mathcal{P}}\mathbf{s}_\theta^{p,c} + (1-q^c)log(1-\frac{1}{N}\sum_{p\in\mathcal{P}}\mathbf{s}_\theta^{p,c})\} + const$$

This is problematic (average distribution of p̂ organ sizes inside log!!)

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (e.g, KL): Partial annotations

primal

dual

$$-\log\alpha = \max_{\beta}\left(\alpha\beta + 1 + \log(-\beta)\right)$$

Stochastic primal-dual gradient

(split terms updated independently)

Prior-aware loss

KL can be expanded

$$-\sum_{c=0}^{|K|}\{q^c log \frac{1}{N}\sum_{p\in\mathcal{P}} \mathbf{s}_\theta^{p,c} + (1-q^c) log(1-\frac{1}{N}\sum_{p\in\mathcal{P}} \mathbf{s}_\theta^{p,c})\} + const$$

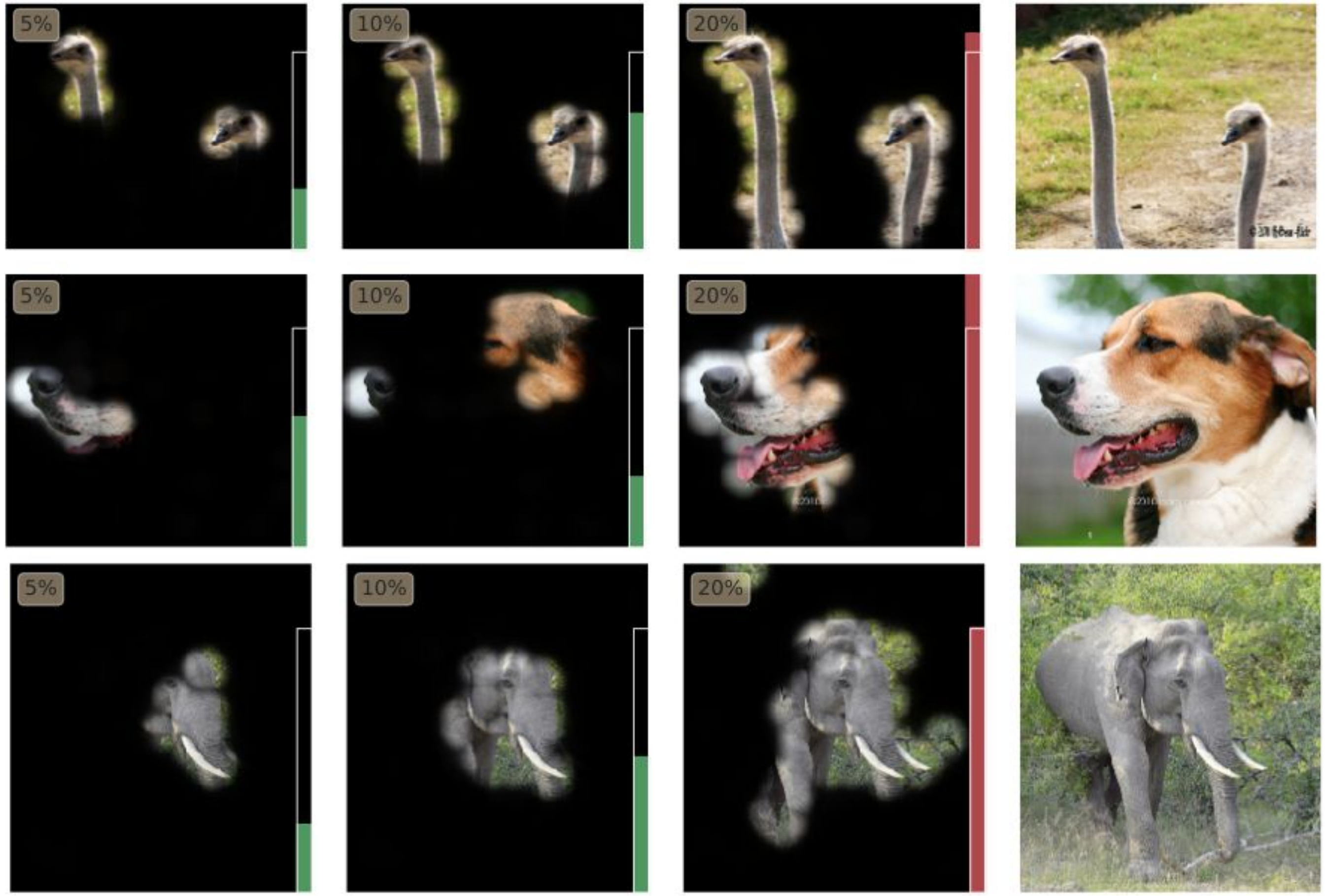
This is problematic (average distribution of p̂ organ sizes inside log!!)

[Zhou et al., Prior-aware Neural Network for Partially-Supervised Multi-Organ Segmentation, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints

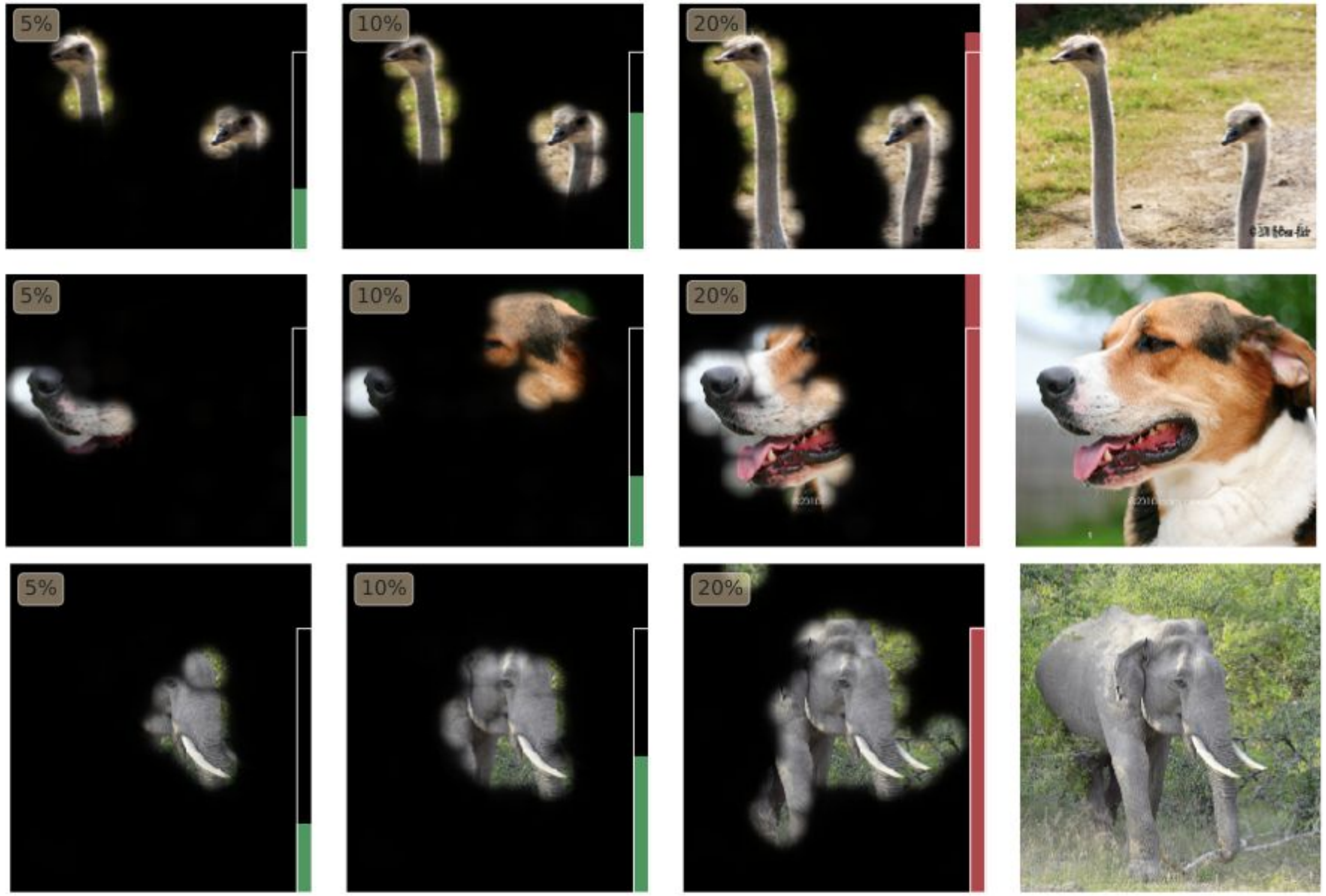
Extremal perturbations



Images from [Fong et al., Understanding Deep Networks via Extremal Perturbations and Smooth Masks, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints



Extremal perturbations

1 - Relax the mask

$$\mathbf{m} \in [0,1]^{|\Omega|}$$

2 - Vectorize and sort **m** (non-decreasing order)

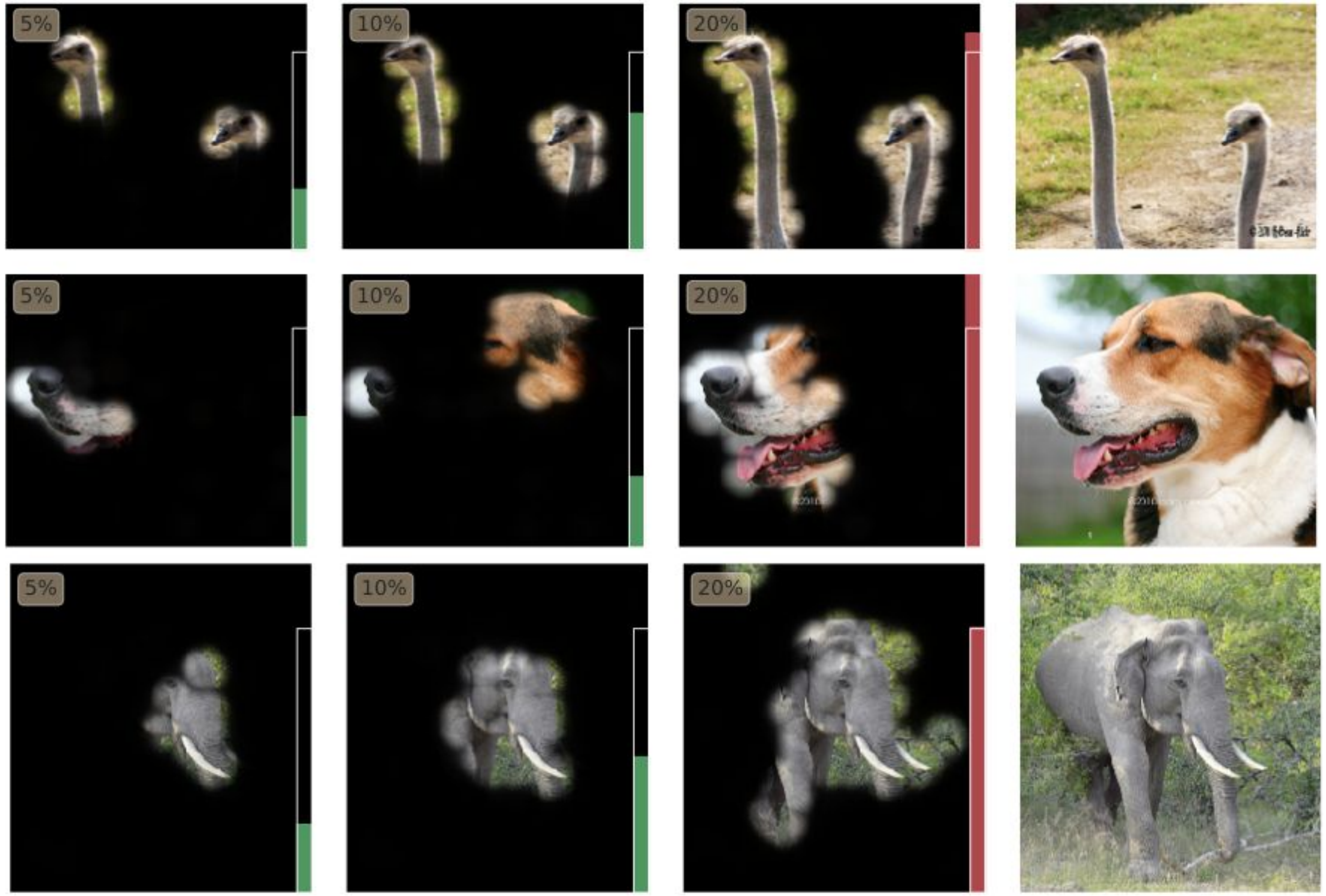
3 - If **m** satisfies the area constraints (a) ***exactly***

$$\mathbf{r}_a \in [0,1]^{|\Omega|} = [\underbrace{0,0,...,0}_{(1-a)|\Omega|},\underbrace{1,1,...,1}_{(a)|\Omega|}]$$

Images from [Fong et al., Understanding Deep Networks via Extremal Perturbations and Smooth Masks, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints



Extremal perturbations

1 - Relax the mask

$$\mathbf{m} \in [0,1]^{|\Omega|}$$

2 - Vectorize and sort **m** (non-decreasing order)

3 - If **m** satisfies the area constraints (a) ***exactly***

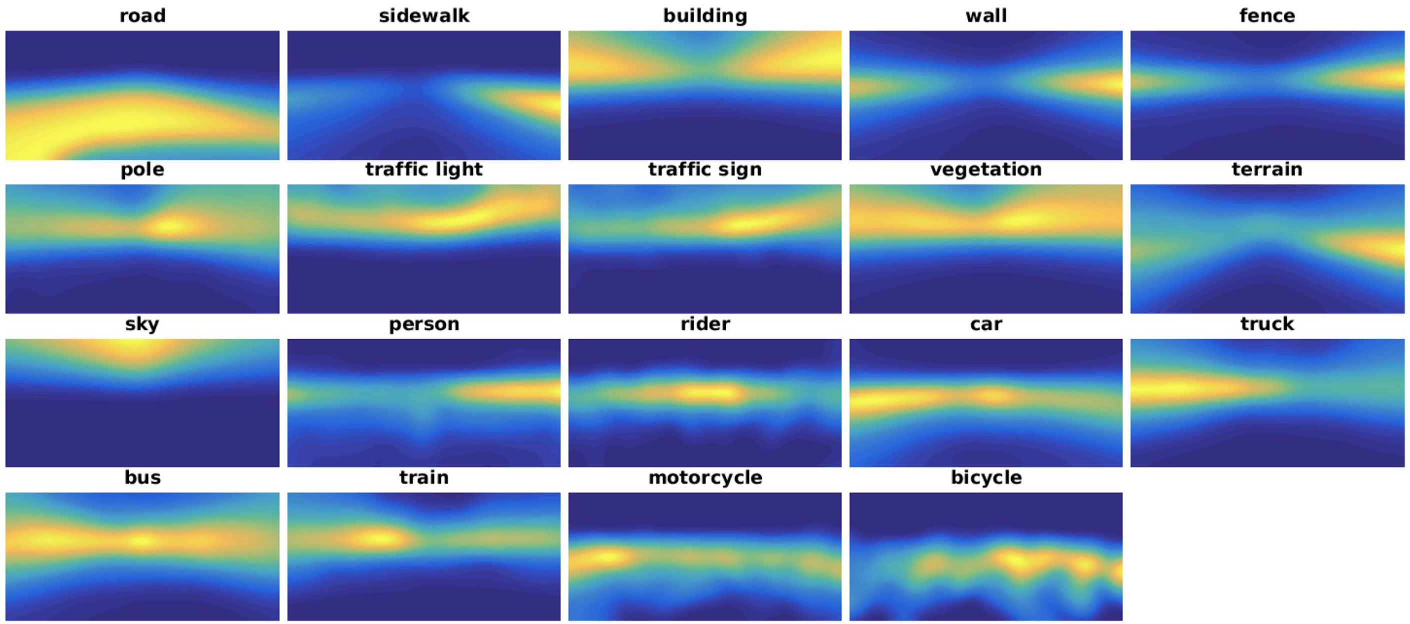
$$\mathbf{r}_a \in [0,1]^{|\Omega|} = [\underbrace{0,0,...,0}_{(1-a)|\Omega|},\underbrace{1,1,...,1}_{(a)|\Omega|}]$$

$$\boxed{R_a(\mathbf{m}) = \|vecsort(\mathbf{m}) - r_a\|^2}$$

Images from [Fong et al., Understanding Deep Networks via Extremal Perturbations and Smooth Masks, ICCV'19]

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)

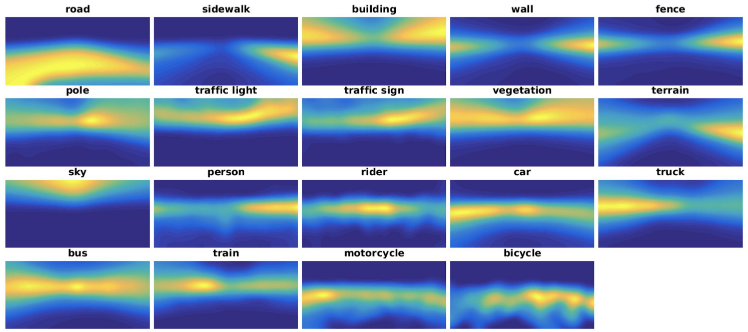


Spatial priors on GTA5

Images from [Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18]

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



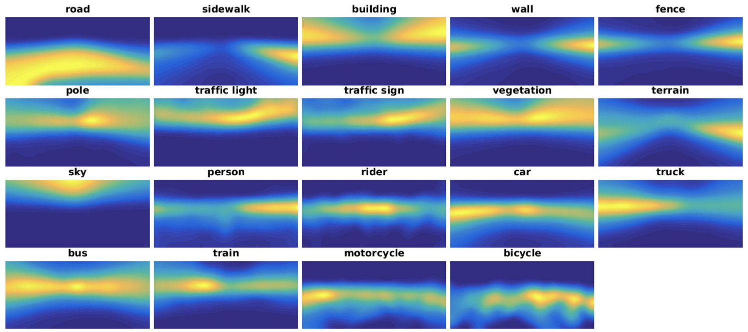
Objective:

$$-\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s_\theta^p) - \sum_{q\in\mathcal{U}}\sum_{c=1}^{K}(\hat{y}^{q,c} log(prior^{q,c} s_\theta^{q,c}) + w_c \hat{y}^{q,c})$$

Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



Proposals

Objective:

$$-\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s_\theta^p) - \sum_{q\in\mathcal{U}}\sum_{c=1}^{K}(\hat{y}^{q,c}log(prior^{q,c}s_\theta^{q,c}) + w_c\hat{y}^{q,c})$$

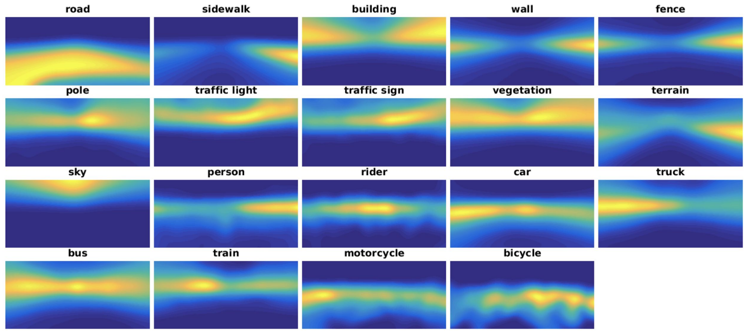
Source images

Target images

Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



Objective:

$$-\sum_{p \in \mathcal{L}} l(\mathbf{y}^p, s_\theta^p) - \sum_{q \in \mathcal{U}} \sum_{c=1}^{K} (\underbrace{\hat{y}^{q,c} log(prior^{q,c} s_\theta^{q,c})} + w_c \hat{y}^{q,c})$$

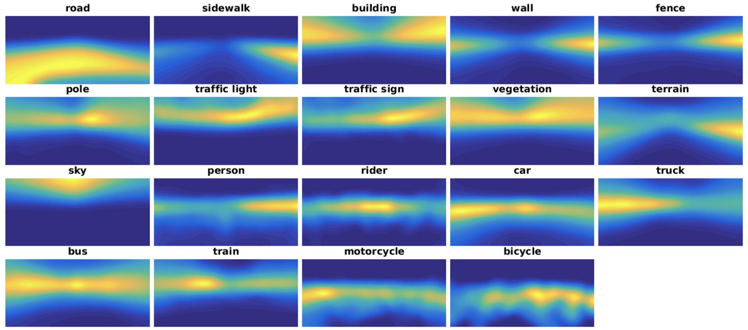
This becomes two KL

$$KL(\hat{y}^{q,c} | prior^{q,c}) \qquad KL(\hat{y}^{q,c} | s_\theta^{q,c})$$

Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)



Objective:

$$-\sum_{p\in\mathcal{L}} l(\mathbf{y}^p, s_\theta^p) - \sum_{q\in\mathcal{U}}\sum_{c=1}^{K}(\hat{y}^{q,c} log(prior^{q,c} s_\theta^{q,c}) + \underbrace{w_c\hat{y}^{q,c}}_{\text{Weights the proposals}})$$

Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18

Constrained optimization (in CNNs)

Equality constraints (at pixel-level)

Spatial prior

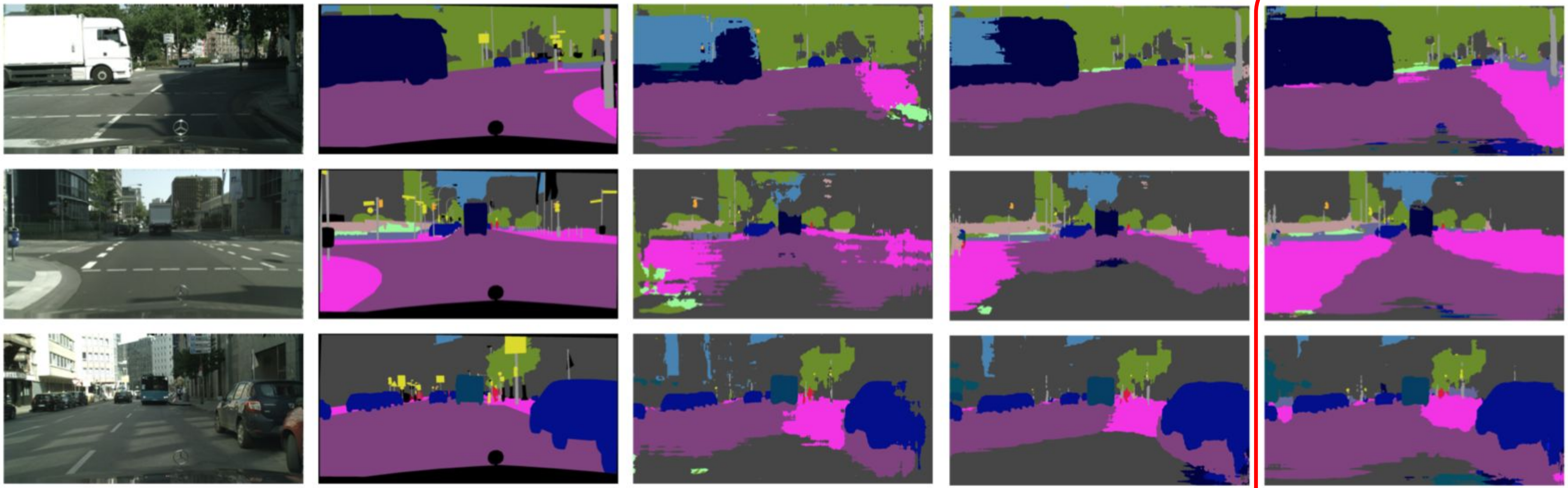


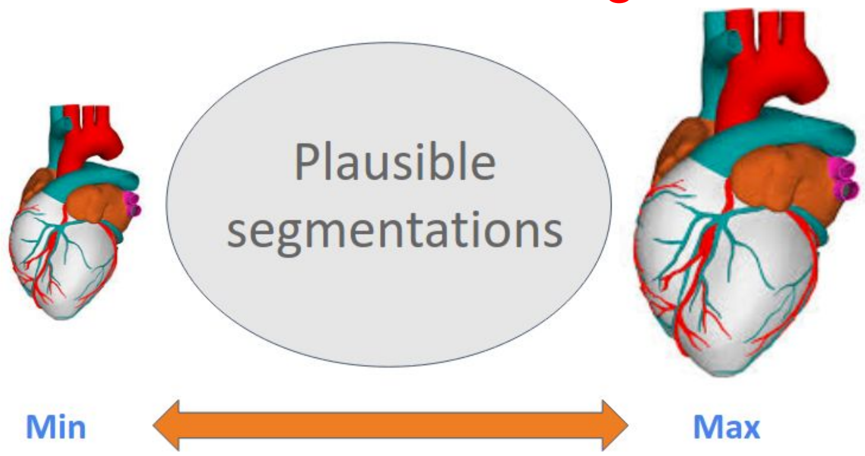
Image | Ground Truth | Only source | CBST | CBST-SP

Images from [Zou et al., Unsupervised Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training, ECCV'18]

Constrained optimization (in CNNs)

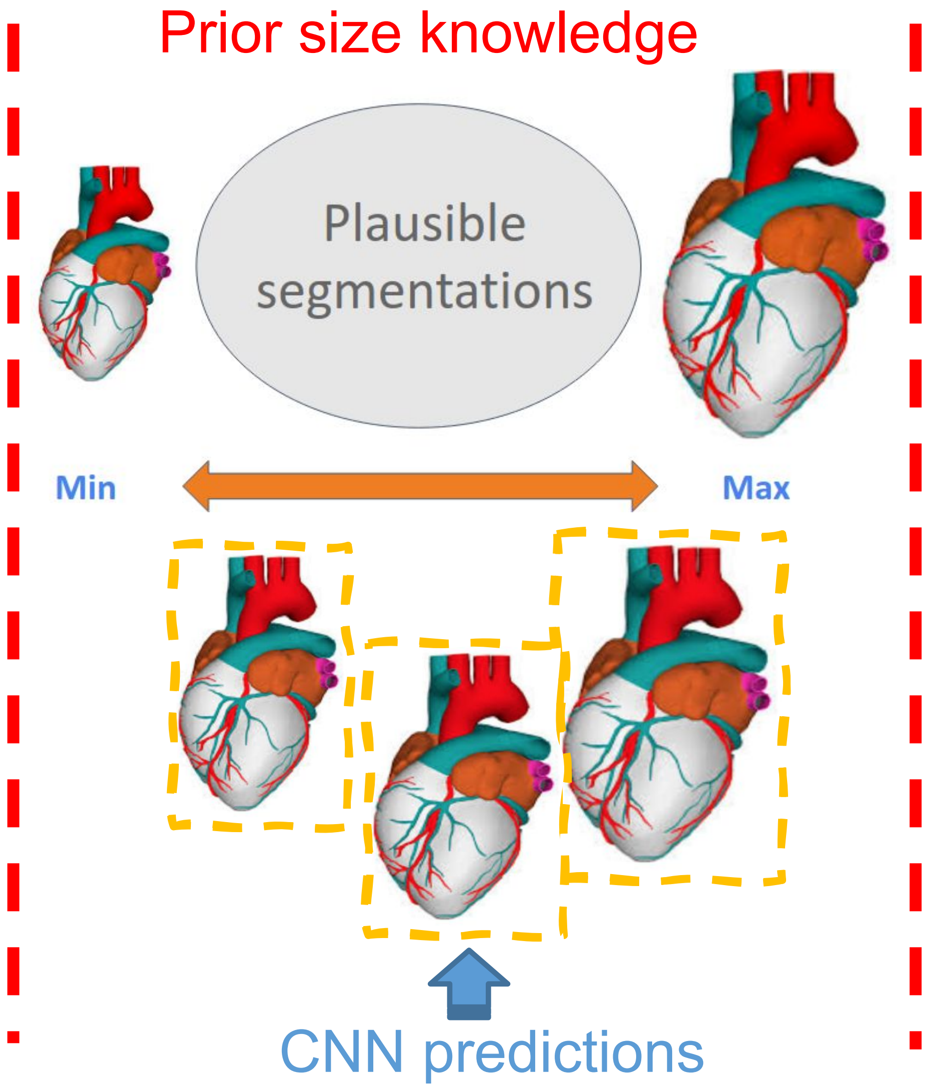
Inequality constraints

Prior size knowledge



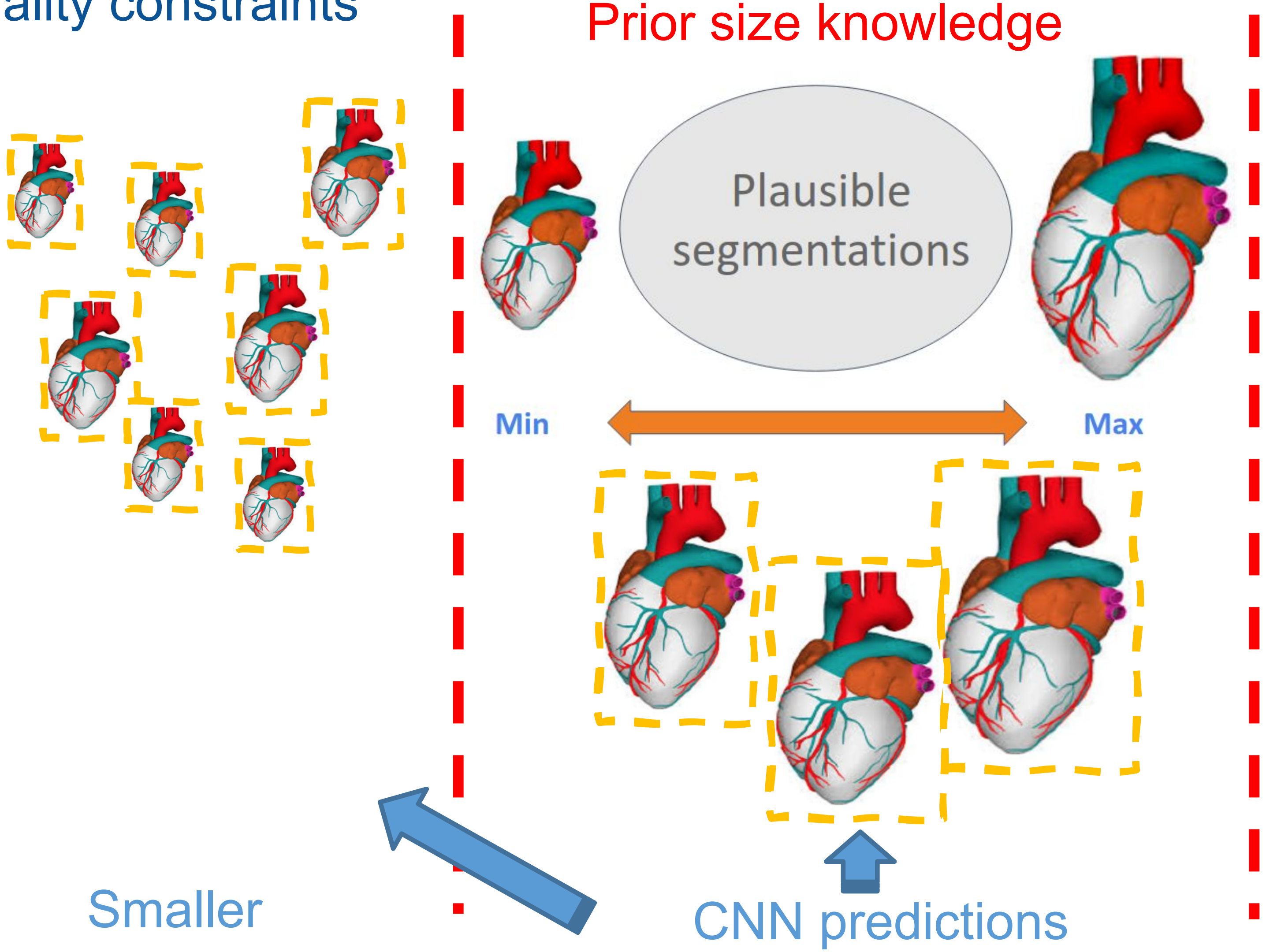
Constrained optimization (in CNNs)

Inequality constraints

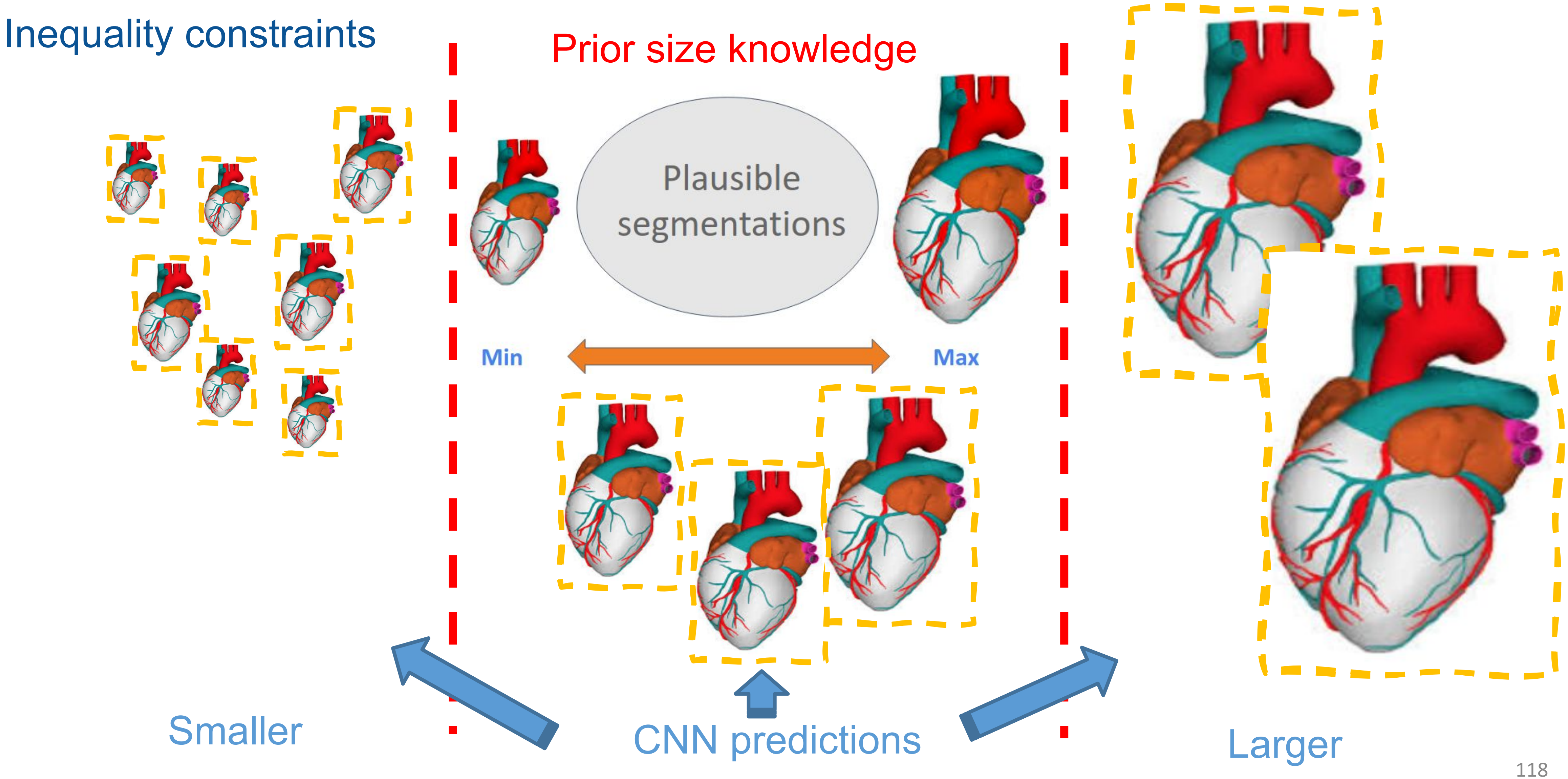


Constrained optimization (in CNNs)

Inequality constraints

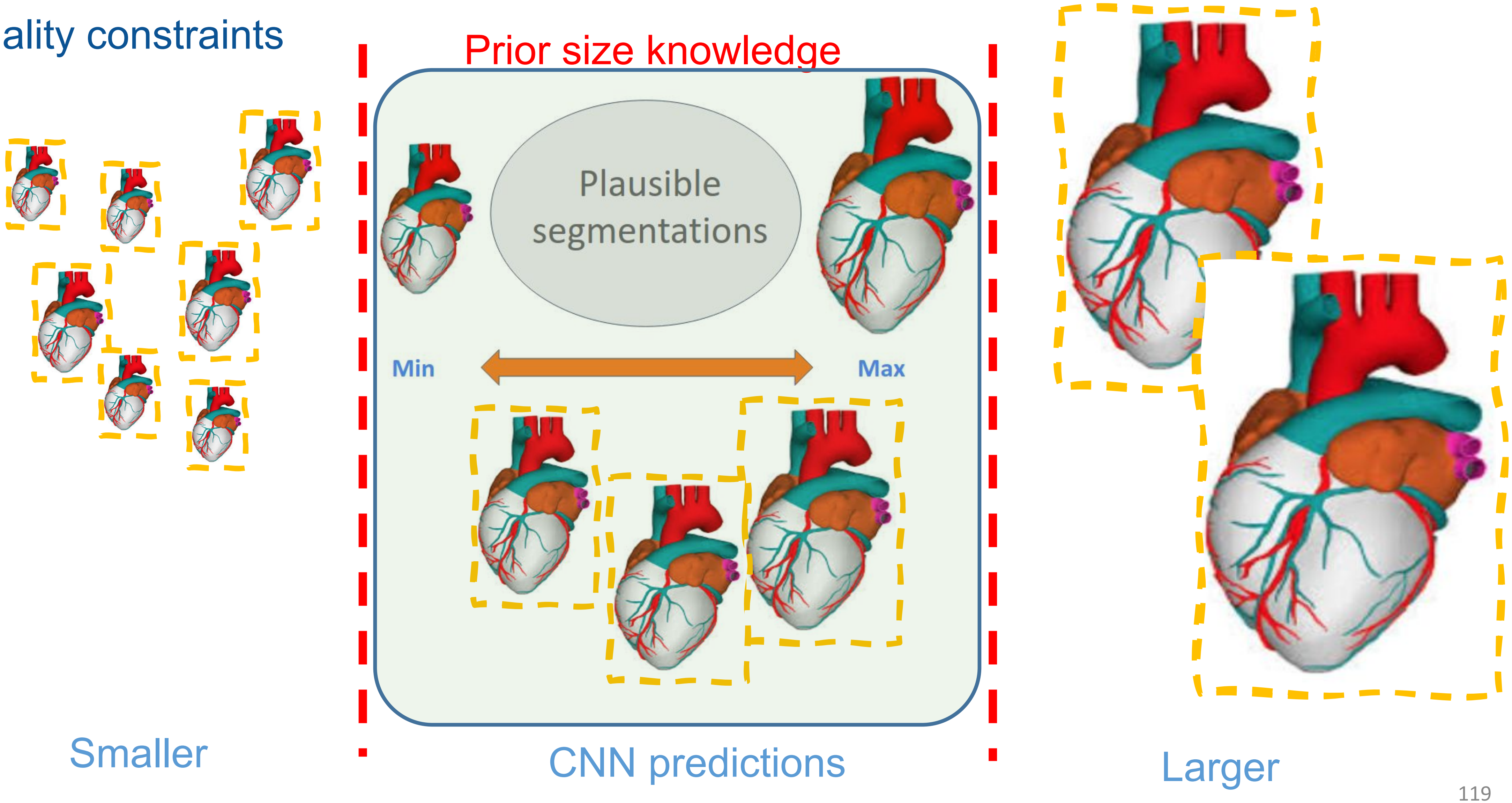


Constrained optimization (in CNNs)

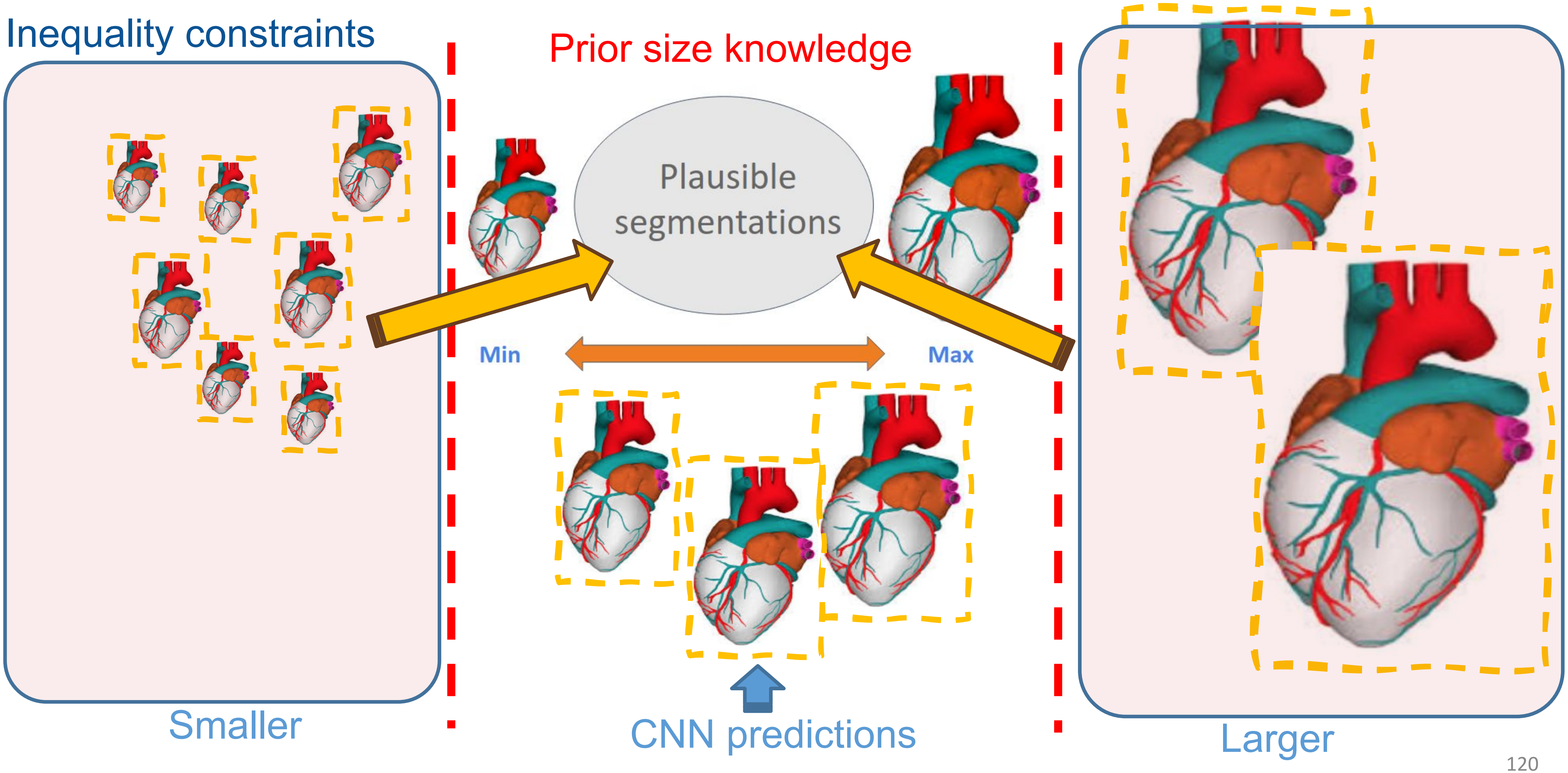


Constrained optimization (in CNNs)

Inequality constraints

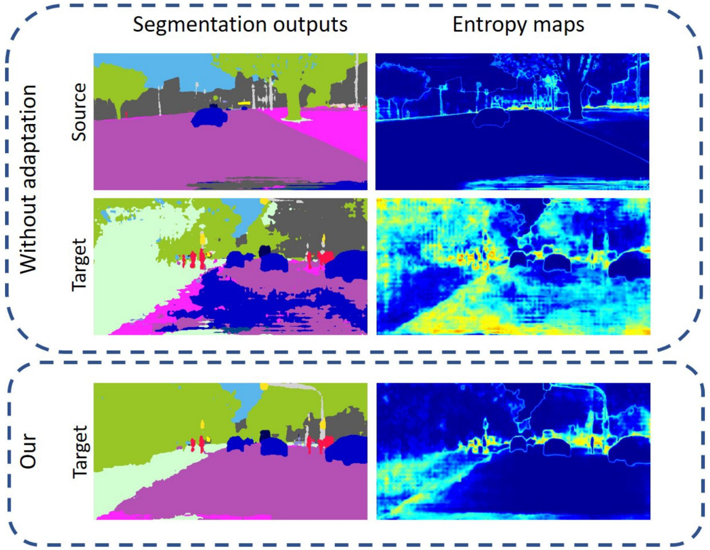


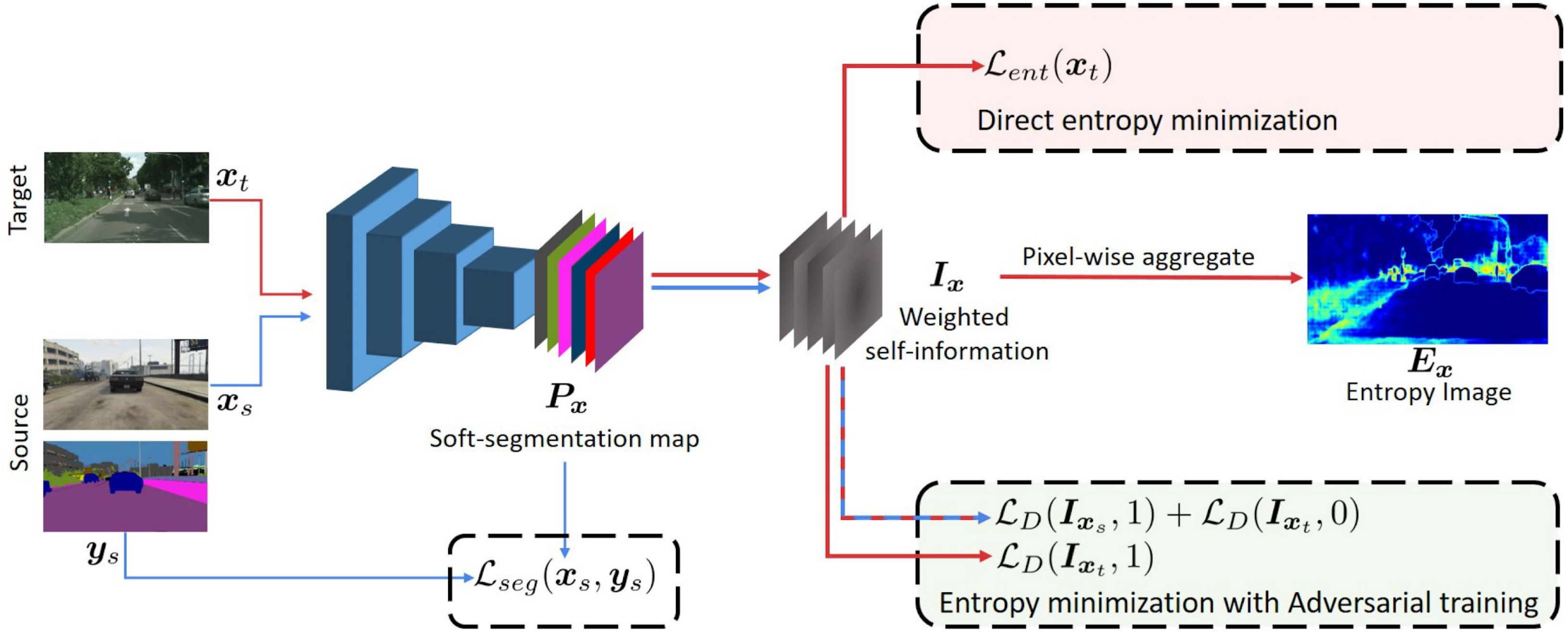
Constrained optimization (in CNNs)



Constrained optimization (in CNNs)

Inequality constraints

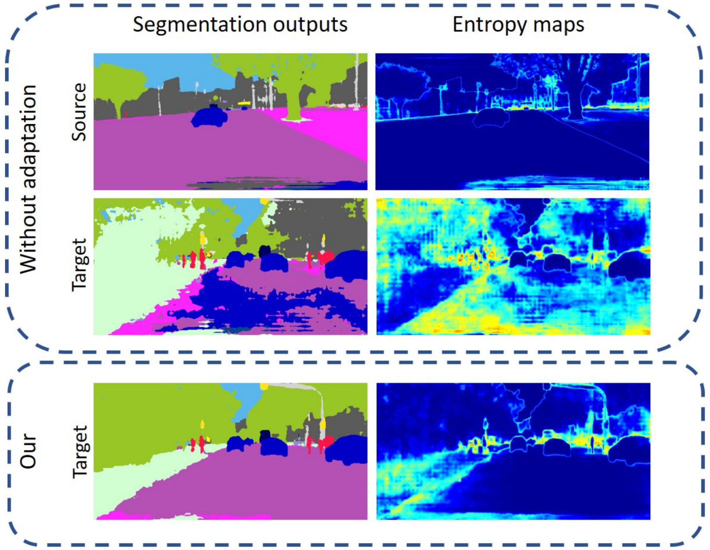


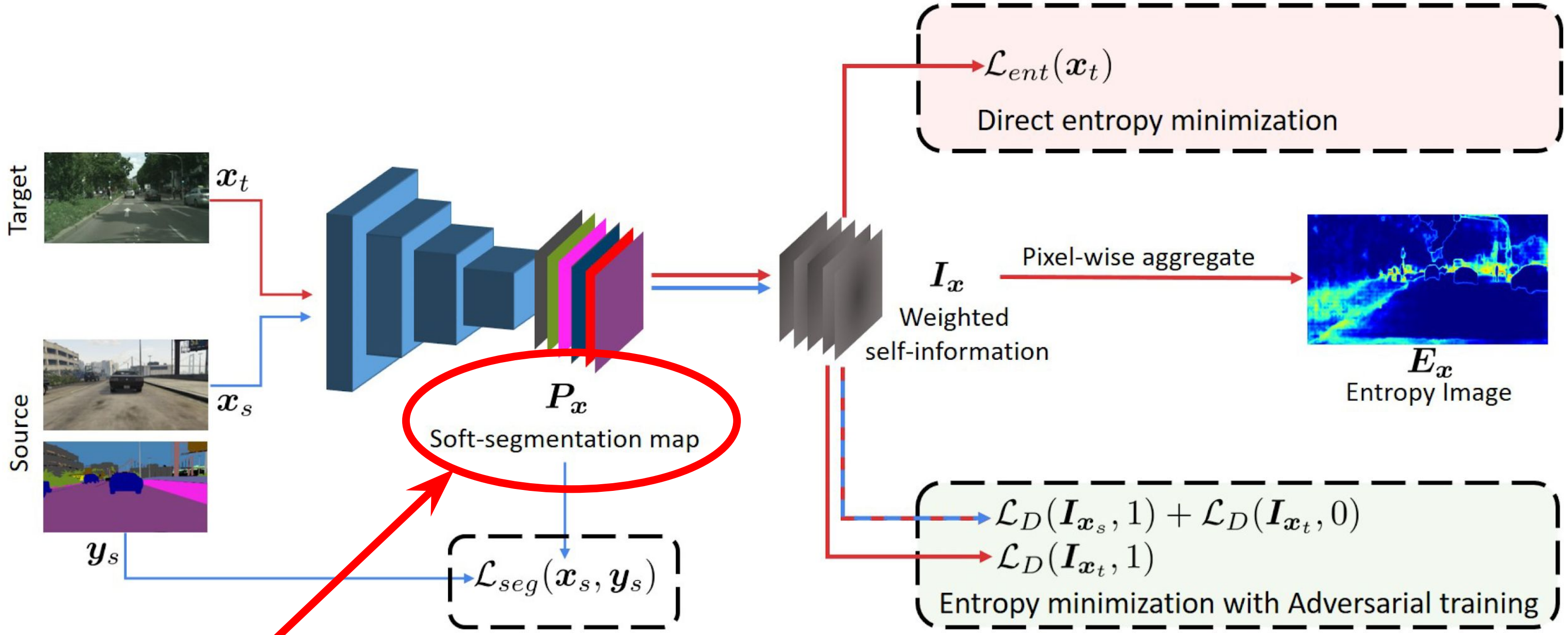


Images from [Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. CVPR'19]

Constrained optimization (in CNNs)

Inequality constraints



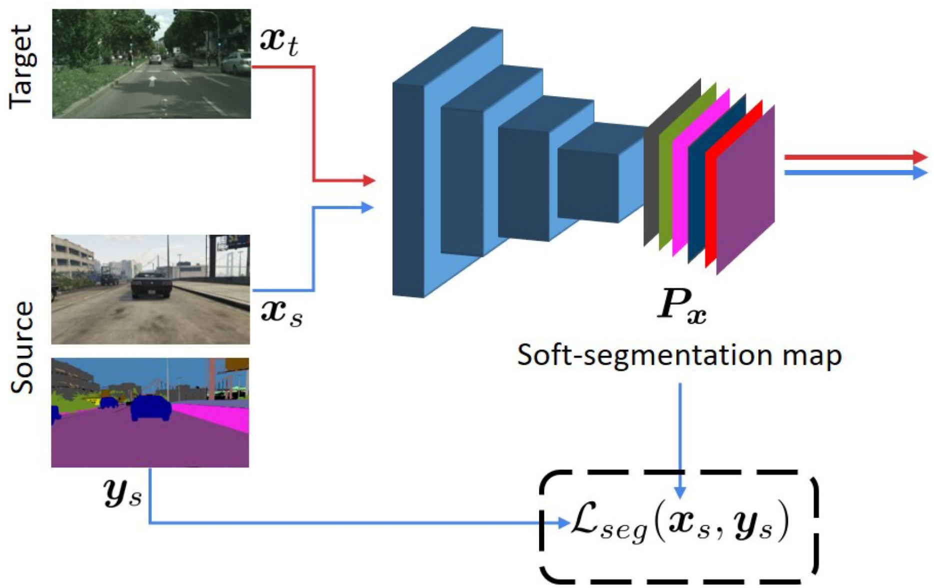


We focus on this now

Images from [Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. CVPR'19]

Constrained optimization (in CNNs)

Inequality constraints



Class-ratio priors

$$\mathcal{L}_{cp}(\boldsymbol{x}_t) = \sum_{c=1}^{C} \max\left(0, \mu \boldsymbol{p}_s^{(c)} - \mathbb{E}_c(\boldsymbol{P}_{\boldsymbol{x}_t}^{(c)})\right)$$

It relaxes the class prior constraint

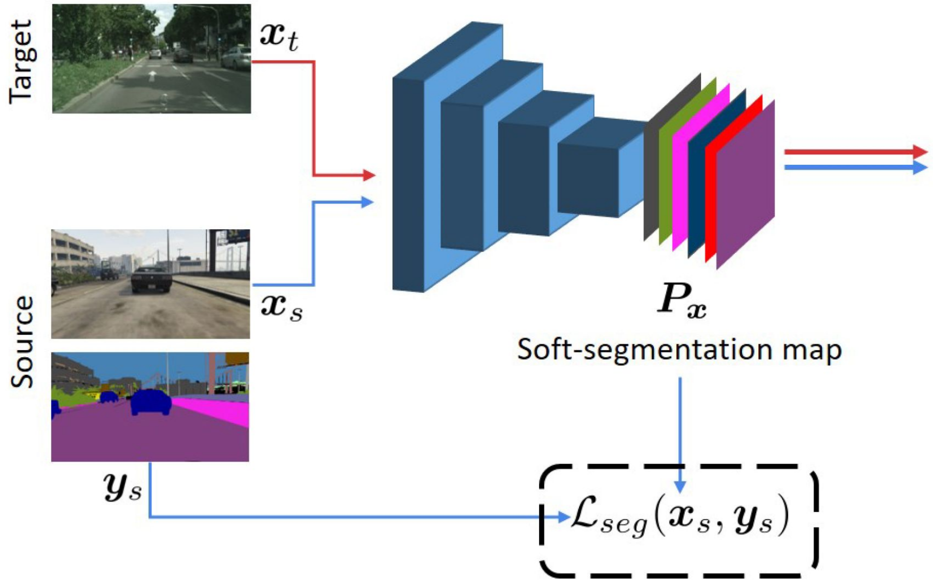
ℓ1-normalized histogram (source)

Estimated size on the prediction

Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. CVPR'19

Constrained optimization (in CNNs)

Inequality constraints



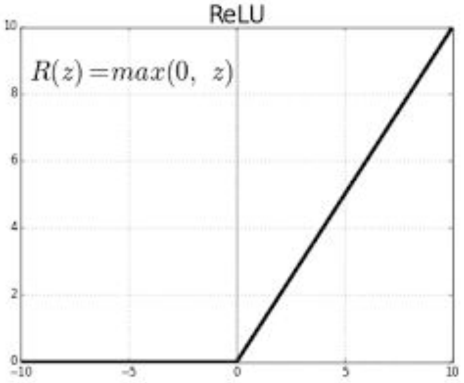
Class-ratio priors

$$\mathcal{L}_{cp}(\boldsymbol{x}_t) = \sum_{c=1}^{C} \max\left(0, \mu \boldsymbol{p}_s^{(c)} - \mathbb{E}_c(\boldsymbol{P}_{\boldsymbol{x}_t}^{(c)})\right)$$

It relaxes the class prior constraint

$\ell 1$-normalized histogram (source)

Estimated size on the prediction



Vu et al., ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. CVPR'19

Constrained optimization (in CNNs)

Inequality constraints

Information is given in the form of image-tags

Suppression $$\sum_{p \in \Omega} s_\theta^{p,c} \leq 0 \quad \forall c \notin C$$

“Person”

[Pathak et al., Constrained Convolutional Neural Networks for Weakly Supervised Segmentation, ICCV’15]

Constrained optimization (in CNNs)

Inequality constraints

Information is given in the form of image-tags

Inclusion (or existence)

$$\sum_{p \in \Omega} s_\theta^{p,c} \geq 1 \quad \forall c \in C$$

“Car”

[Pathak et al., Constrained Convolutional Neural Networks for Weakly Supervised Segmentation, ICCV’15]

Constrained optimization (in CNNs)

Inequality constraints

Information is given in the form of image-tags

Target Size
a > 1

$$\sum_{p \in \Omega} s_\theta^{p,c} \geq a \quad \forall c \in C$$

"Car"

[Pathak et al., Constrained Convolutional Neural Networks for Weakly Supervised Segmentation, ICCV'15]

Constrained optimization (in CNNs)

How we can benefit from this in the medical domain?

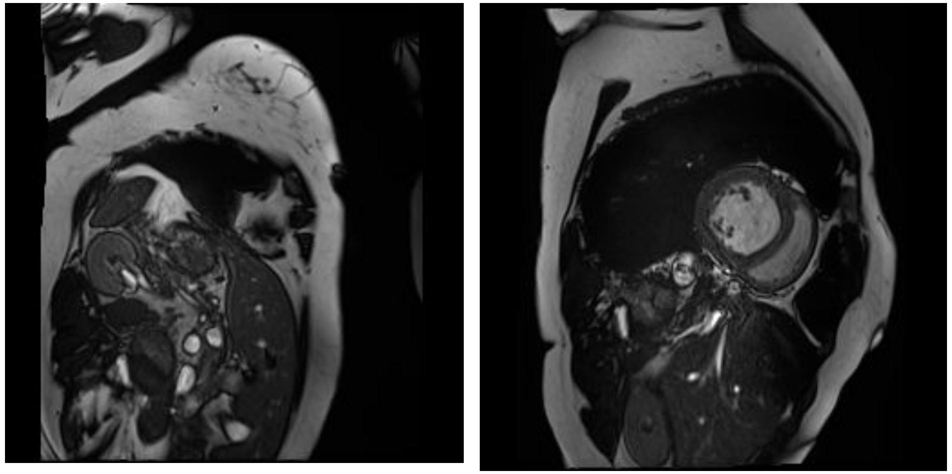


Image-tag information

$$\sum_{p \in \Omega} s_\theta^{p,c} \leq 0$$

For negative image tags

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

How we can benefit from this in the medical domain?

No cavity   Cavity

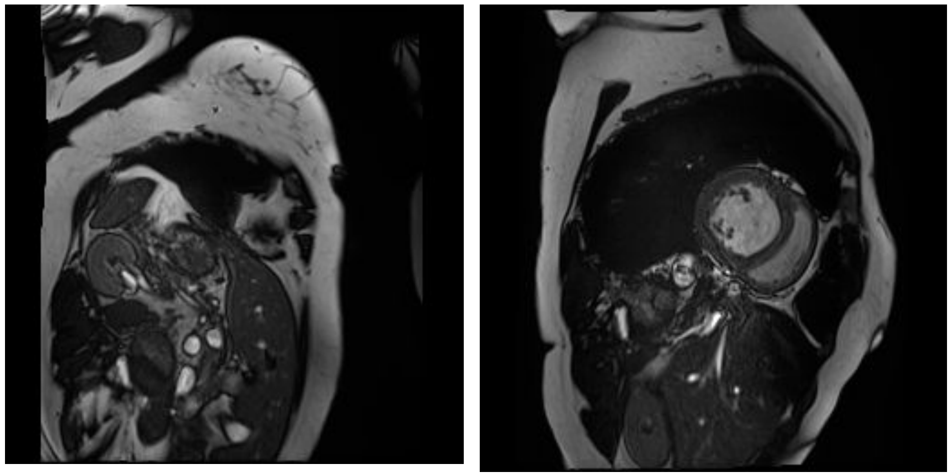
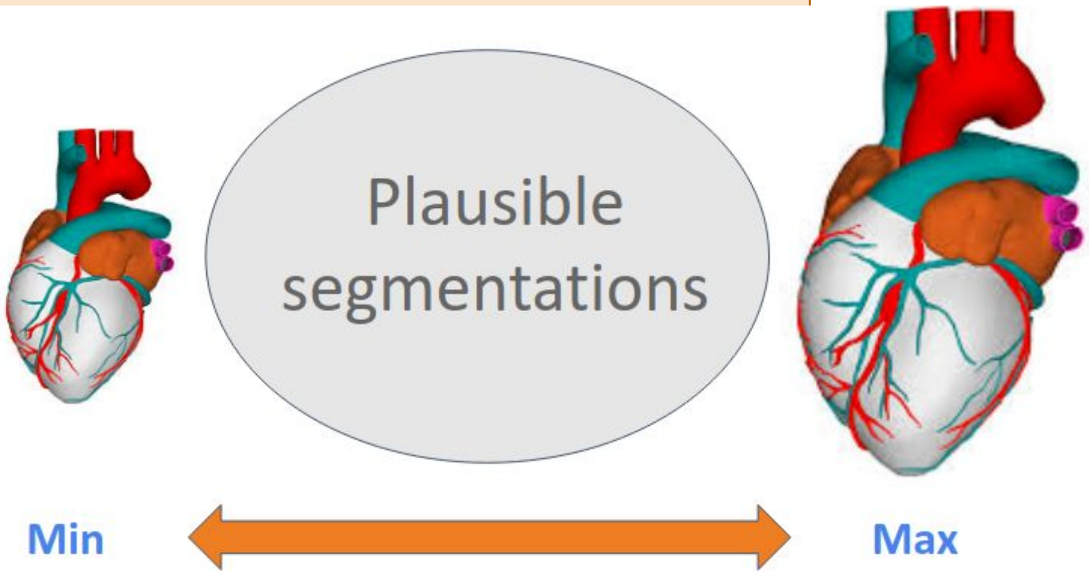


Image-tag information

$$\sum_{p \in \Omega} s_{\theta}^{p,c} \leq 0$$

For negative image tags



Size information

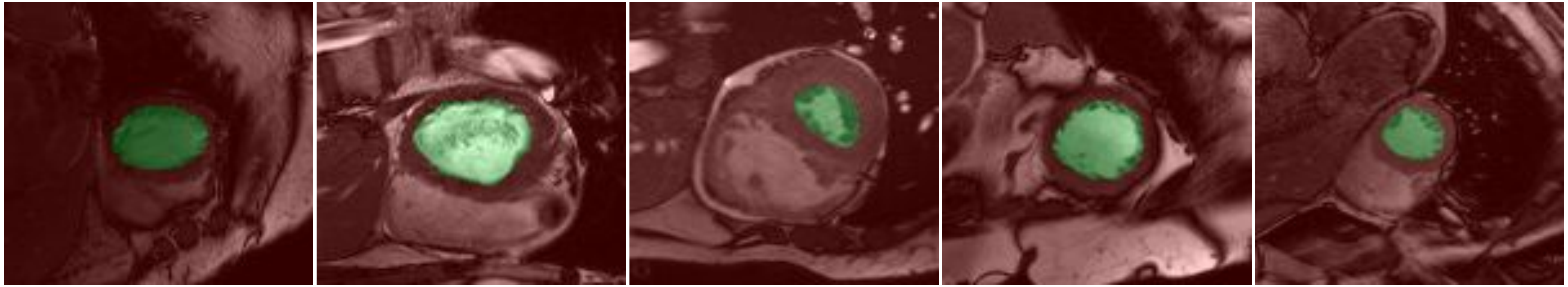
$$min \leq \sum_{p \in \Omega} s_{\theta}^{p,c} \leq max$$

For positive image tags

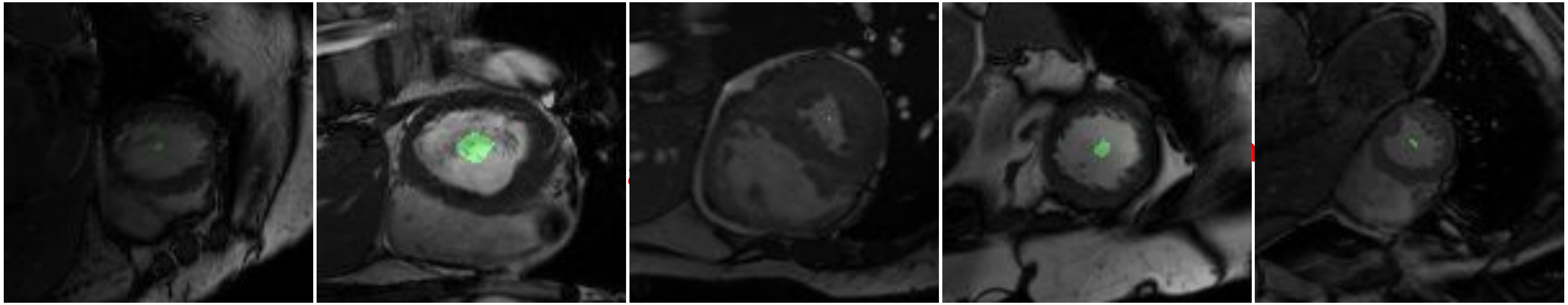
[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)



Full annotations



Partial annotations for cross-entropy

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Objective

$$\min_{\boldsymbol{\theta}} \mathcal{H}(S) \quad \text{s.t} \quad a \leq \sum_{p \in \Omega} s_{\theta}^{p,c} \leq b \quad \Rightarrow \quad \mathcal{H}(S) + \lambda \mathcal{C}\left(V_S\right)$$

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Objective

$$\min_{\boldsymbol{\theta}} \mathcal{H}(S) \quad \text{s.t} \quad a \leq \sum_{p \in \Omega} s_{\theta}^{p,c} \leq b \quad \Rightarrow \quad \mathcal{H}(S) + \lambda \mathcal{C}\left(V_S\right)$$

$$\mathcal{H}(S) = -\sum_{p \in \mathcal{L}} log(s_{\theta}^{p})$$

On annotated pixels

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Objective

$$\min_{\boldsymbol{\theta}} \mathcal{H}(S) \quad \text{s.t} \quad a \leq \sum_{p \in \Omega} s_\theta^{p,c} \leq b \quad \Rightarrow \quad \mathcal{H}(S) + \lambda \mathcal{C}(V_S)$$

$$\mathcal{H}(S) = -\sum_{p \in \mathcal{L}} log(s_\theta^p)$$

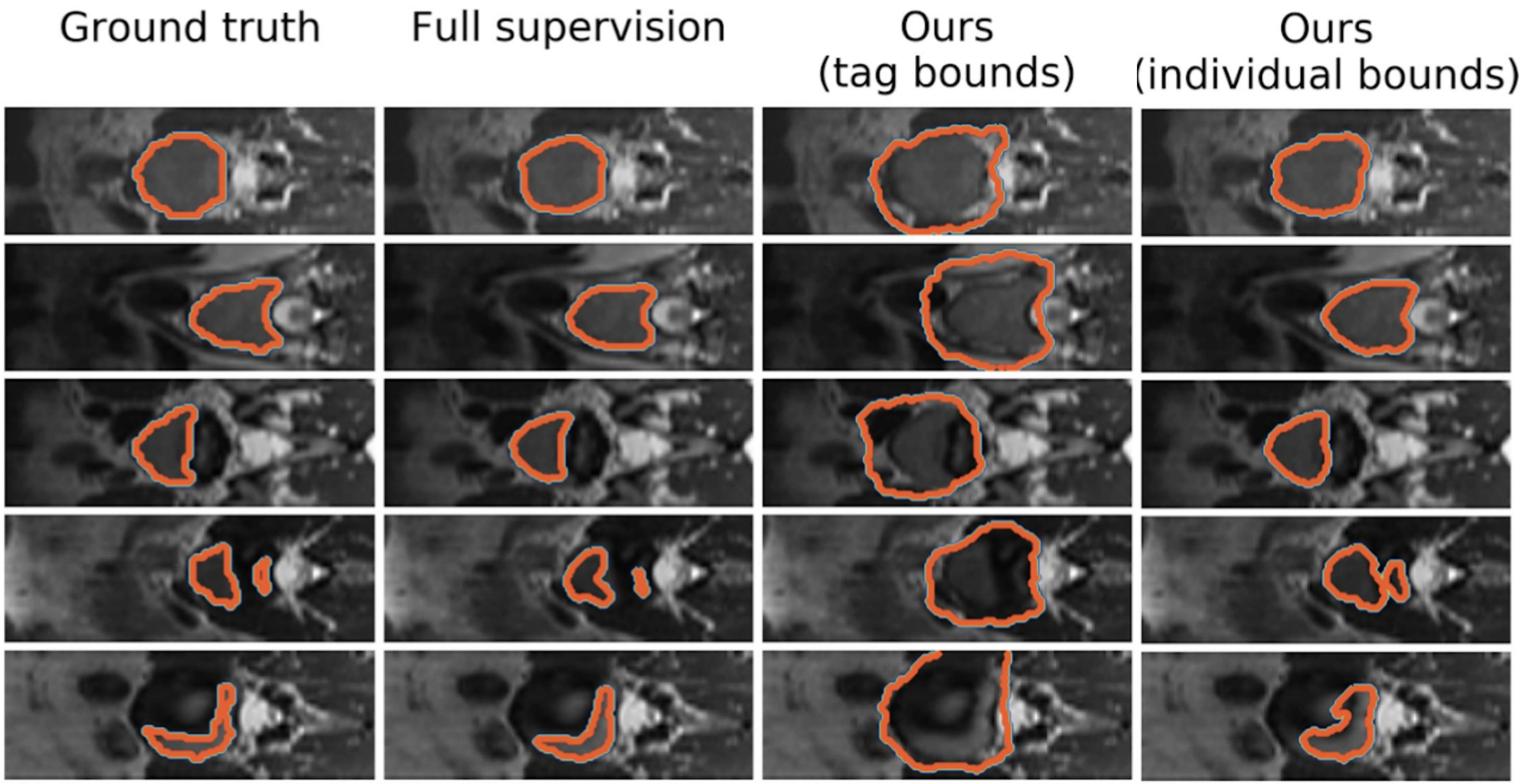
On annotated pixels

$$\mathcal{C}(V_S) = \begin{cases} (V_S - a)^2, & \text{if } V_S < a \\ (V_S - b)^2, & \text{if } V_S > b \\ 0, & \text{otherwise} \end{cases}$$

[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

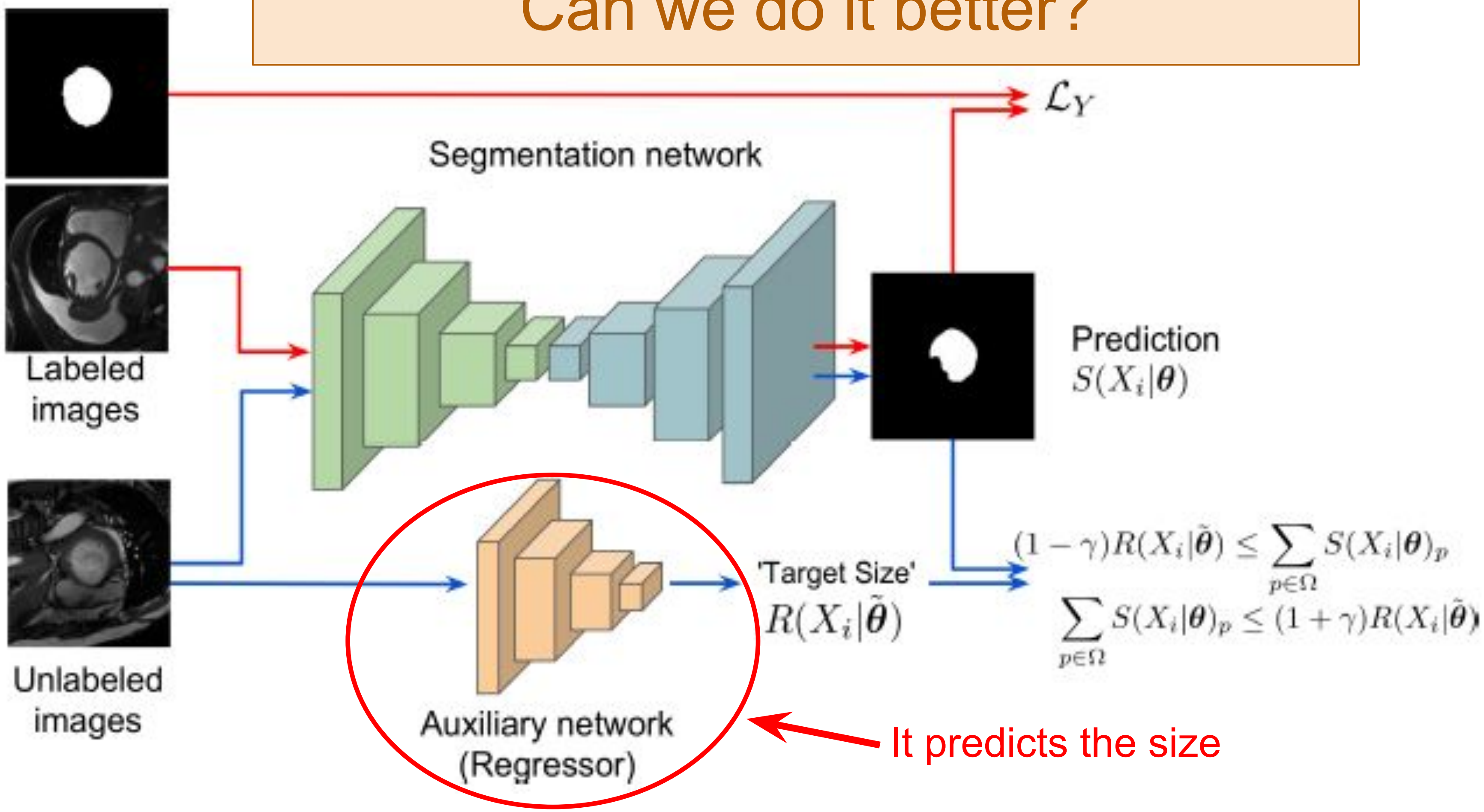


[Kervadec et al., Constrained-CNN losses for weakly supervised segmentation. MedIA 2019]

Constrained optimization (in CNNs)

Inequality constraints (e.g, L2 penalty)

Can we do it better?



[Kervadec et al., Curriculum semi-supervised segmentation. MICCAI'19]

Take-home message

- Imposing constraints helps weakly-supervised segmentation learning by restricting plausible segmentations on unlabeled images

- Few constraints have been explored under low-labeled data regime

- Room for improvement (many opportunities)

Thank you!